UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE AUTOMAÇÃO E SISTEMAS
CURSO DE GRADUAÇÃO EM ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Pedro de Oliveira Lima Xavier

Modelos de categorização de imóveis e de predição de faturamento e taxa de ocupação de imóveis

Florianópolis
2022

Pedro de Oliveira Lima Xavier

Modelos de categorização de imóveis e de predição de faturamento e taxa de ocupação de imóveis

Relatório final da disciplina DAS5511 (Projeto de Fim de Curso) como Trabalho de Conclusão do Curso de Graduação em Engenharia de Controle e Automação da Universidade Federal de Santa Catarina em Florianópolis.
Orientador: Prof. Marcelo de Lellis Costa de Oliveira, Dr.
Supervisor: Bruno Benetti, Eng.

Florianópolis
2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Xavier, Pedro de Oliveira Lima
Modelos de categorização de imóveis e de predição de faturamento e taxa de ocupação de imóveis / Pedro de Oliveira Lima Xavier ; orientador, Marcelo de Lellis Costa de Oliveira, coorientador, Bruno Benetti, 2022.
57 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Controle e Automação, Florianópolis, 2022.

Inclui referências.

1. Engenharia de Controle e Automação. 2. Modelos Classificatórios. 3. Modelos Preditivos. 4. Aluguel de imóveis por temporada. 5. Airbnb. I. Oliveira, Marcelo de Lellis Costa de . II. Benetti, Bruno. III. Universidade Federal de Santa Catarina. Graduação em Engenharia de Controle e Automação. IV. Título.

Pedro de Oliveira Lima Xavier

Modelos de categorização de imóveis e de predição de faturamento e taxa de ocupação de imóveis

Esta monografia foi julgada no contexto da disciplina DAS5511 (Projeto de Fim de Curso) e aprovada em sua forma final pelo Curso de Graduação em Engenharia de Controle e Automação

Florianópolis, 13 de Dezembro de 2022.

Prof. Hector Bessa Silveira, Dr.
Coordenador do Curso

Banca Examinadora:

Prof. Marcelo de Lellis Costa de Oliveira, Dr.
Orientador
UFSC/CTC/DAS

Bruno Benetti, Eng.
Supervisor
Empresa Seazone

Prof. Eduardo Hulse, Dr.
Avaliador
UFSC/CTC

Prof. Eduardo Camponogara, Dr.
Presidente da Banca
UFSC/CTC/DAS

Este trabalho é dedicado à minha família e aos meus colegas de curso e de trabalho que me apoiaram ao longo de toda a minha graduação. Dedico também aos professores e servidores da UFSC que foram fundamentais na minha formação como profissional e como indivíduo.

AGRADECIMENTOS

Gostaria de agradecer aos meus companheiros de trabalho Artur Brito, Lucas Abel, Bárbara Farah, Francisco Burigo, Isabella Guerreiro, Márcio Fazolin, Augusto Hideki, Gabriel Probst, Davi Marques, aos meus supervisores André Crescenzo e Bruno Benetti e ao meu professor orientador Marcelo de Lellis por terem contribuído imensamente na produção deste trabalho, além de constantemente me motivarem a estar sempre melhorando e elevando o nível das minhas entregas e contribuições e de me apoiarem nos momentos em que enfrentei problemas e dificuldades. Também gostaria de agradecer especialmente a minha mãe, quem esteve comigo em todos os momentos da minha vida e principalmente durante a minha formação como Engenheiro de Controle e Automação pela UFSC.

RESUMO

No mercado de aluguel por temporada, as dinâmicas de precificação e reservas dos imóveis são muito mais aceleradas em comparação ao mercado de aluguel tradicional. Por isso, requer um entendimento maior das particularidades de cada imóvel para que sejam realizadas ações que melhorem a performance do imóvel ao longo do ano e que acompanhem esta dinâmica mais rápida deste mercado. Informações como qualidade do imóvel, ocupação ao longo do tempo e faturamento ao longo do tempo são de extrema importância neste cenário. Assim, este trabalho propõe o desenvolvimento e implementação de três modelos baseados em dados, utilizando a base de dados obtida através do Airbnb. Entre eles: um modelo de categorização dos imóveis, um modelo de predição da taxa de ocupação dos imóveis ao longo do ano e um modelo de predição de faturamento dos imóveis ao longo do ano. Estes modelos serão utilizados pela empresa Seazone para ampliar sua capacidade de análise e decisão em relação aos imóveis administrados pela empresa.

Palavras-chave: Dados. Modelos. Predição. Classificação. Imóveis. Airbnb. Aluguel. Computação em nuvem. Computação distribuída.

ABSTRACT

On vacation rental market, the pricing and reservation dynamics are far more accelerated than compared to the traditional rental market. For that reason, it requires a bigger understanding of the particularities of each property, so that the proper action to improve a listing performance is taken over the year. Information such as a listing quality, its occupation over the year and the revenue it makes over time are of extreme importance in this market. That being said, this paper proposes the development and implementation of three data-based models utilizing the database obtained through Airbnb, such as a categorization model for Airbnb listing, a prediction model for the occupancy rate of those listings over the year and a revenue prediction model for those listings over the year. These models will then be utilized by the company Seazone, in order to increase its capability of analyzing and decision making over the listings administrated by the company.

Keywords: Data. Models. Prediction. Classification. Listing. Airbnb. Rental. Cloud computing. Distributed computing.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em linhas gerais, o trabalho aqui proposto trata da elaboração de modelos matemáticos preditivos e classificatórios para imóveis anunciados no Airbnb, os quais são objeto de análise de negócio para a empresa Seazone Serviços Ltda.

O escopo do projeto se estende a todos os imóveis do Brasil anunciados na plataforma. Cada anúncio passa por um processo de *web scrapping*, no qual todas as informações referentes àquele imóvel são obtidas e armazenadas no banco de dados da empresa. Estes dados brutos são tratados e limpos para que possam assim ser utilizados em modelos que forneçam informações mais profundas, precisas e valiosas sobre os imóveis.

Assim, o foco deste trabalho se dará nesta etapa de elaboração de modelos, podendo dividir o projeto em duas partes: o modelo de categorização dos imóveis e os modelos de predição de faturamento e taxa de ocupação dos imóveis. Cada parte possui motivações, objetivos, metodologias e propostas específicas, as quais serão descritas neste capítulo introdutório.

1.1 MOTIVAÇÃO

Modelo de categorização de imóveis

Anteriormente a este trabalho, os imóveis disponíveis no banco de dados da empresa eram categorizados com base na análise das fotos de cada anúncio. Isto é, uma das tarefas do cotidiano da empresa consistia em acessar os anúncios individualmente, interpretar a qualidade do imóvel com base nas fotos ali disponíveis e definir uma categoria para aquele imóvel. Este processo era lento, sem um critério universal bem estabelecido, pouco escalável e de difícil atualização, além de ser necessário alocar uma pessoa treinada que realizasse esta categorização. Ou seja, estudar e implementar formas de automatizar este processo poderia gerar resultados muito positivos em todos esses pontos.

Além disso, a categorização dos imóveis possui duas principais utilidades para a Seazone:

- Precificação dos imóveis.
- Clusterização dos imóveis em grupos similares para análise.

Nos primórdios da empresa, a precificação dos imóveis - que é um dos processos mais fundamentais da empresa - era feita de forma manual. O processo já era feito com base na análise de dados, no entanto, para realizar a precificação, uma pessoa era responsável por analisar os preços de imóveis concorrentes aos da Seazone (com base em valores e informações disponíveis apenas em planilhas) e com base nesta

análise e na análise de sazonalidade, alterar os valores de cada uma das diárias da semana de cada um dos imóveis da Seazone, toda semana.

Este processo já foi automatizado internamente, com uma ferramenta batizada de Sirius. Em resumo, esta ferramenta elabora regras de precificação com base em uma série de critérios e dados e as aplica de forma automatizada em todos os imóveis da Seazone, precificando-os. Um dos *inputs* mais fundamentais desta ferramenta é justamente a categoria dos imóveis, já que cada categoria de imóvel terá um conjunto de regras de negócio e de precificação próprias.

Por fim, a categorização dos imóveis também é utilizada para que a empresa possa acompanhar o desempenho dos imóveis separando-os em grupos similares, o que é de grande utilidade, já que imóveis de diferentes qualidades e características possuem diferentes metas e comportamentos ao longo do ano e devem ser considerados separadamente.

Modelos de predição de faturamento e ocupação de imóveis

Já para a motivação dos modelos de predição, temos que há duas principais utilidades desprendidas das informações de faturamento e ocupação dos imóveis para a Seazone:

- Precificação dos imóveis.
- Captação de novos clientes e investimentos.

Ter de antemão uma predição de faturamento e ocupação para cada imóvel, para cada preço de diária definido e para cada data do ano, nos permite traçar uma curva de faturamento (preço da diária x taxa de ocupação) e maximizar esta curva. Ou seja, com estes modelos, sabemos exatamente que preço colocar para cada imóvel de forma a maximizar o faturamento deste imóvel para qualquer data. Também sabemos o quanto um imóvel será alugado em uma dada época do ano com base no preço pelo qual este imóvel está sendo ofertado e o quanto esta ocupação pode variar com a variação do preço da diária, o que nos permite “controlar” a ocupação variando o preço. Caso um imóvel esteja sendo alugado abaixo do esperado podemos abaixar o valor da diária, aumentando sua ocupação e, analogamente, caso um imóvel esteja sendo alugado acima do esperado podemos aumentar o valor da diária, sempre em busca de um ponto máximo de faturamento.

Além disso, a Seazone está caminhando para um modelo de franquias, expandindo para novas localidades e, por isso, está constantemente buscando novos proprietários de imóveis e novos possíveis investidores dispostos a comprar ou construir novos imóveis em parceria com a empresa. Para isso, é muito importante fornecer uma predição de faturamento para um imóvel em potencial para que o proprietário

possa avaliar o que a empresa está oferecendo e escolher deixar o seu imóvel em nossa custódia ou para um investidor que está buscando comprar ou construir imóveis em parceria com a Seazone possa saber o que esperar de retorno do seu possível investimento. Isso traz credibilidade e confiança para a empresa e contribui para a captação e retenção de novos clientes.

Por fim, há ainda uma última - porém não menos importante - motivação que se dá no elemento *Big Data* do projeto. Já foram feitas no passado tentativas de criar estes modelos na Seazone. No entanto, conforme o banco de dados da empresa expandiu com o tempo, tornou-se necessário estudar novas tecnologias capazes de lidar com uma quantidade de dados massiva, já que eventualmente as tecnologias até então utilizadas não eram mais capazes de manipular e processar a quantidade de dados necessária, causando até *crashes* de computadores e inviabilizando a execução de diversos modelos baseados em dados fundamentais para o cotidiano da empresa.

1.2 OBJETIVOS

Modelo de categorização de imóveis

O objetivo do modelo de categorização consiste na elaboração de um algoritmo baseado em dados capaz de categorizar, de forma coerente e universal, todos os imóveis disponíveis no banco de dados da empresa em cinco categorias, em ordem crescente de qualidade do imóvel: SIM (simples), JR (júnior), SUP (superior), TOP (top), MASTER (master). Assim, o *output* esperado do modelo é que para cada imóvel deve ser atribuído uma categoria, necessariamente (internamente chamamos de *strata*).

*nModelos de predição de faturamento e ocupação de imóveis Já para os modelos de predição de faturamento e ocupação, temos que o objetivo consiste em desenvolver um algoritmo que seja capaz de separar os dados disponíveis em diversos cenários de análise - levando em consideração localização, sazonalidade (mês), número de quartos, tipo do imóvel, feriados, fins de semana, a própria categoria do imóvel obtida pelo modelo de categorização, entre outros - e atribuir para cada cenário, uma predição de faturamento mensal e a probabilidade de ocupação (0-100%) para o conjunto de imóveis daquele cenário.

1.3 A EMPRESA - SEAZONE SERVIÇOS LTDA

A Seazone é uma empresa de aluguel de imóveis por temporada atualmente sediada em Florianópolis, mas que já está expandindo para diversas outras cidades do Brasil, com o objetivo de crescer para o país todo. Os clientes principais da empresa são proprietários de imóveis que buscam multiplicar o faturamento de seus imóveis a partir das tecnologias e serviços que oferecemos e também se isentar das preocupações e dos problemas que administrar um imóvel pode gerar.

Por isso, a Seazone oferece um serviço de gestão de imóveis de ponta a ponta. Fazemos a manutenção e limpeza dos imóveis, cuidamos de *check-in* e *checkout* dos locatários, remodelamos o anúncio do imóvel, tiramos fotos profissionais do imóvel e fazemos a precificação dinâmica do imóvel ao longo do ano, sempre buscando otimizar a ocupação e faturamento do imóvel a partir do preço da diária.

Fazemos isso com eficácia e consistência por sermos acima de tudo uma empresa de tecnologia totalmente baseada em dados. Possuímos os dados de todos os imóveis do Airbnb do Brasil, os quais utilizamos para diversas análises e construção de modelos - como os propostos neste trabalho - que nos ajudam a garantir que estamos precificando e administrando os nossos imóveis da melhor forma possível, maximizando a ocupação dos imóveis ao longo do ano todo e maximizando o faturamento dos proprietários.

1.4 REVISÃO BIBLIOGRÁFICA

1.4.1 Hipótese de Mercado Eficiente

O ponto de partida que orientou a revisão bibliográfica realizada neste trabalho foi a hipótese de Mercado Eficiente. Esta hipótese foi inicialmente proposta no âmbito do Mercado Financeiro e afirma que o preço de um ativo reflete todas as informações possíveis disponíveis sobre a instituição emissora, como descrito no artigo de Mussa *et al.* (2004). Ou seja, qualquer informação positiva sobre a instituição emissora disponível no mercado terá uma contribuição positiva no preço do ativo e vice-versa. Esta ideia foi extrapolada no âmbito deste trabalho, para o mercado imobiliário.

Levantou-se a hipótese de que o preço de um imóvel reflete todas as informações disponíveis sobre aquele imóvel. Isso permitiria pensar em um modelo que monitore preço como variável manipulada principal, extraindo diversos outros parâmetros como variáveis de entrada, como: tamanho, localização, número de quartos, qualidade, mobília, *amenities*, entre outros. Assim, seria possível *clusterizar* preços de imóveis em grupos similares de análise e tanto definir uma categoria para estes grupos de imóveis (modelo de categorização de imóveis) quanto calcular predições de faturamento e ocupação para estes grupos (modelos de predição de faturamento e ocupação de imóveis).

No entanto, há inúmeras maneiras de se *clusterizar* preço, por isso foi realizada a seguinte revisão bibliográfica para escolher e entender os métodos estatísticos e matemáticos que viriam a ser utilizados neste trabalho.

1.4.2 Clusterização de preço

Tanto para o modelo de categorização de imóveis, quanto para o modelo de predição de faturamento, a capacidade de clusterizar preço com base em variáveis

determinantes dos imóveis é fundamental. Para este fim, inicialmente buscou-se artigos de classificação e predições já previamente enviesados no contexto deste trabalho, ou seja, voltados para o mercado imobiliário de temporada. Encontrou-se, primeiramente, o artigo deKirkos (2022), o qual foi um bom ponto de partida, rendendo uma noção geral de que possíveis caminhos poderiam ser tomados. Um dos desafios iniciais do problema de clusterização de preço para o contexto deste trabalho é a escolha das variáveis de entrada do modelo. Neste artigo, o autor sugere alguns possíveis parâmetros determinantes na distinção e agrupamento de imóveis, como: número de quartos, avaliações do anúncio, localização do imóvel, capacidade de hóspedes, comodidades, entre outros.

Além das seleções das variáveis, o autor também sugere algumas técnicas possíveis para realizar a predição de faturamento e ocupação de imóveis do Airbnb. Entre elas, o autor implementou cinco algoritmos distintos. Para o caso deste trabalho, as técnicas de árvores de decisão (*decision trees* - DT), redes neurais (*Multilayer perceptron neural network* - MLP) e florestas aleatórias (*Random Forests* - RF) seriam alternativas possíveis para os problemas de clusterização de preço (modelo de faturamento) e classificação de imóveis (modelo de categorização). Já técnicas como regressões logísticas (*logistic regression* - LR) e de máquina de vetores de suporte (*Support vector machines* - SVM) poderiam ser utilizadas para predizer a taxa de ocupação dos imóveis (modelo de ocupação). O autor ainda ressalta que são métodos especialmente bons para predizer a porcentagem de um evento binário (*booleano*), como é o caso de se uma diária foi reservada no Airbnb ou não.

Além desses métodos, o autor também sugere modelos heurísticos para estimar faturamento e ocupação de imóveis ou até para estimar a quantidade de reservas que um imóvel terá em um ano. Todos os algoritmos foram estudados mais a fundo, avaliando outros exemplos de aplicações implementados em outros artigos.

Embora o estudo aprofundado destas técnicas tenha sido proveitoso e provido alguns caminhos possíveis para este trabalho, os artigos que melhor se encaixaram com a ideia de clusterização de preço para categorizar imóveis e predizer faturamento foram os artigos sobre Regressão Quantílica, que se mostraram ser exatamente o que estava sendo buscado e estes serão tratados com maior profundidade nesta seção.

Inicialmente, o primeiro artigo visitado que mostrou grande potencial foi o de Geraci (2019). Nele, o autor aplica o algoritmo de regressão quantílica em alguns casos de estudo distintos, utilizando diferentes bases de dados. Destaca-se dois em que o resultado mostrou-se muito próximo do objetivo desejado para este trabalho: no primeiro caso, o autor utiliza a regressão quantílica para avaliar o comportamento da concentração de um medicamento após ser injetado em pacientes ao longo do tempo, separando este comportamento em faixas de concentração, como mostrado na figura a seguir:

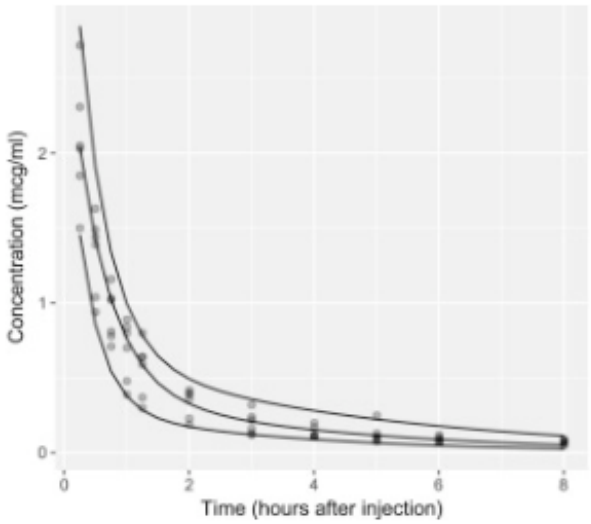


Figura 1 – Faixas de concentração de medicamento ao longo do tempo.
Fonte: (GERACI, 2019)

Estas faixas são determinadas a partir dos quantis definidos no modelo de regressão quantílica. Neste caso, por exemplo, o autor utilizou os quantis de 10%, 50% e 90% como parâmetros para *clusterizar* os dados de concentração.

Outro caso estudado neste artigo buscou avaliar a dinâmica do peso das folhas de pés de soja ao longo do tempo, também separando-os em faixas de peso:

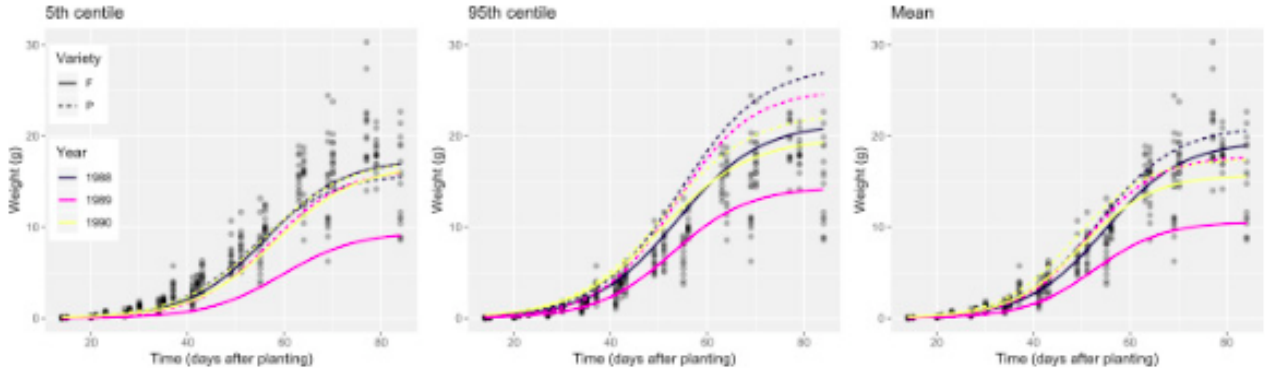


Figura 2 – Faixas de peso das folhas de pés de soja ao longo do tempo.
Fonte: (GERACI, 2019)

Mais uma vez, estas faixas são determinadas pelos quantis definidos no modelo de regressão quantílica. No gráfico a esquerda, temos os dados *clusterizados* no quantil de 5%, enquanto que no gráfico ao meio temos o quantil de 95% e no gráfico da direita temos a média (a mediana poderia ser obtida com o quantil de 50%).

Estes resultados mostraram-se muito promissores para o caso de estudo deste trabalho, já que seria possível separar o preço dos imóveis em faixas, de forma que cada faixa poderia ser uma categoria de um imóvel, alocando cada imóvel de cada faixa a uma categoria. Além disso, seria possível calcular quantis intermediários de cada faixa de forma a obter um preço médio dos imóveis de uma categoria, o que

combinado ao modelo de taxa de ocupação poderia fornecer um resultado satisfatório de predição de faturamento. Estes conceitos são explicados mais a fundo no capítulo 3 e os resultados gerados, os quais foram muito similares aos mostrados nas figuras anteriores, estão explícitos no capítulo 4.

1.4.3 Predição da taxa de ocupação dos imóveis

Como mencionado anteriormente, o artigo de Kirkos (2022)já forneceu dois possíveis caminhos para a predição da taxa de ocupação dos imóveis: utilizar a regressão logística ou uma máquina de vetores de suporte, ambas funcionais para predizer a porcentagem de ocorrência de um evento tipicamente *booleano*, como é o caso de uma reserva de uma diária no Airbnb. Neste artigo, o autor foi capaz de predizer a taxa de ocupação de diferentes grupos de imóveis com precisão de aproximadamente 80% utilizando um modelo de regressão logística, o que seria satisfatório para o caso deste trabalho, sendo possível propor melhorias para melhorar este resultado. Mesmo assim, realizou-se um estudo mais aprofundado destes métodos buscando ampliar a validação e o entendimento destes, avaliando-os em outros artigos com aplicações similares.

Inicialmente, optou-se por investigar a regressão logística, a qual parecia ser vastamente utilizada para predições deste tipo e de implementação mais consolidada, difundida e prática.

Em Dasguptaa e Deb (2007)o autor implementou um modelo de regressão logística para prever a probabilidade da ocorrência de chuvas de curto prazo (próximas 12 horas), com base na análise de eventos binários de ocorrência de chuvas por dia (choveu ou não choveu em um determinado dia) em um determinado período.

Outra fonte fundamental tanto para o entendimento e validação da regressão logística como solução, quanto para sua implementação, foi o artigo de Hansen (2021) publicado na página da Universidade da California em São Diego (UCSD). Neste artigo o autor utiliza da regressão logística para calcular a probabilidade de ocorrência de diversos eventos binários analisados e discorre sobre sua implementação. Entre eles, o autor é capaz de predizer a probabilidade de sobrevivência de passageiros do Titanic separadas por gênero, idade e tipo da passagem comprada (1ª classe e 2ª classe). Com uma complexidade ainda maior, analisando ainda mais variáveis de entrada, o autor também realizou a predição da probabilidade de um projeto em uma escola pública nos Estados Unidos receber doações da *Donors Choose*, uma instituição sem fins lucrativos dos Estados Unidos que permite que indivíduos doem diretamente para projetos deste tipo. No modelo de regressão logística construído, o autor utilizou um conjunto de variáveis de entrada mais complexo em comparação ao caso estudo de caso do Titanic. Ao todo 14 variáveis de entrada foram utilizadas, entre elas: localização do colégio, características do professor responsável, grau de escolaridade na qual o

projeto seria implementado, custo, número de estudantes, entre outros.

Em todos os casos discutidos anteriormente, os modelos de regressão logística se mostraram satisfatórios para predizer a probabilidade de ocorrência de um evento binário com base em uma série de variáveis, o que seria exatamente os requisitos para a predição da taxa de ocupação dos imóveis desejada neste trabalho.

METODOLOGIA E ESTRUTURA DO DOCUMENTO

A metodologia aplicada na execução e validação deste trabalho consta, inicialmente, com a definição do objetivo e do escopo do trabalho. Em seguida, a partir dos objetivos e do escopo definido, é realizada a revisão bibliográfica para escolher os algoritmos, técnicas e tecnologias a serem utilizadas na execução do trabalho. O objetivo e escopo do trabalho e a revisão bibliográfica são descritas neste primeiro capítulo introdutório.

As tecnologias e ferramentas utilizadas são apresentadas no capítulo 2. Entre elas, destaca-se o uso do *framework* para computação distribuída: PySpark, utilizado em conjunto com a tecnologia de *clusters* na nuvem, disponibilizadas pela Amazon Web Services (AWS - EMR), as quais foram fundamentais para a manipulação e processamento dos dados em larga escala.

Já os métodos matemáticos e estatísticos escolhidos, juntamente com a construção e detalhamento dos algoritmos são descritos de forma mais profunda no capítulo 3. É também necessário definir o conjunto de dados a ser utilizado no cálculo dos modelos, as variáveis de entrada e saída esperadas e os procedimentos de limpeza, tratamento e pré-processamento dos dados para os modelos. Todos estes pontos também são tratados neste capítulo.

Em seguida, é preciso definir e realizar os testes para avaliação do desempenho dos algoritmos implementados, os quais são apresentados no capítulo 4 **??**. Vale ressaltar que a metodologia empregada na definição dos testes baseia-se no conceito de MVP's (*Minimum Viable Product*), na qual as entregas são bem definidas a partir de um protótipo funcional minimalista para validação de uma etapa do projeto.

As etapas de validação e discussão dos resultados, análise e definição de métricas e possíveis otimizações do modelo configuram as últimas etapas do projeto e estão presentes no capítulo 4.

Por fim, a conclusão do trabalho e considerações finais estão retratadas no sexto e último capítulo 5.

2 AMBIENTE DE TRABALHO E TECNOLOGIAS UTILIZADAS

2.1 PYSPARK

Como mencionado anteriormente, a questão de trabalhar com tecnologias comumente utilizadas para trabalhos de *Big Data* a fim de garantir escalabilidade e eficiência no processamento de dados em larga escala é fundamental para este projeto. Como estamos trabalhando com tabelas contendo todas as informações de todas as diárias de todos os imóveis do Airbnb do Brasil, estamos lidando com dados na ordem de bilhões de registros, o que resulta em tabelas com centenas de *Gigabytes* de memória.

Por isso, a primeira tecnologia a ser aqui descrita é o *framework open source* Apache Spark, mais especificamente o PySpark, uma API que permite utilizar o Apache Spark em algoritmos na linguagem de programação Python.

O PySpark é um *framework* que conta com uma série de bibliotecas e ferramentas para executar trabalhos em dados de forma distribuída, particionando os dados e distribuindo-os em diversos nós em um *cluster* (conjunto de computadores). Ou seja, diversos problemas de escalabilidade que enfrentamos na empresa ao trabalhar com uma grande quantidade de dados, processando-os em uma máquina só de forma centralizada, agora podem ser resolvidos distribuindo-os em múltiplas máquinas que processam os dados de forma paralela.

A arquitetura de um trabalho distribuído em PySpark está representada nas figuras 11 e 4.

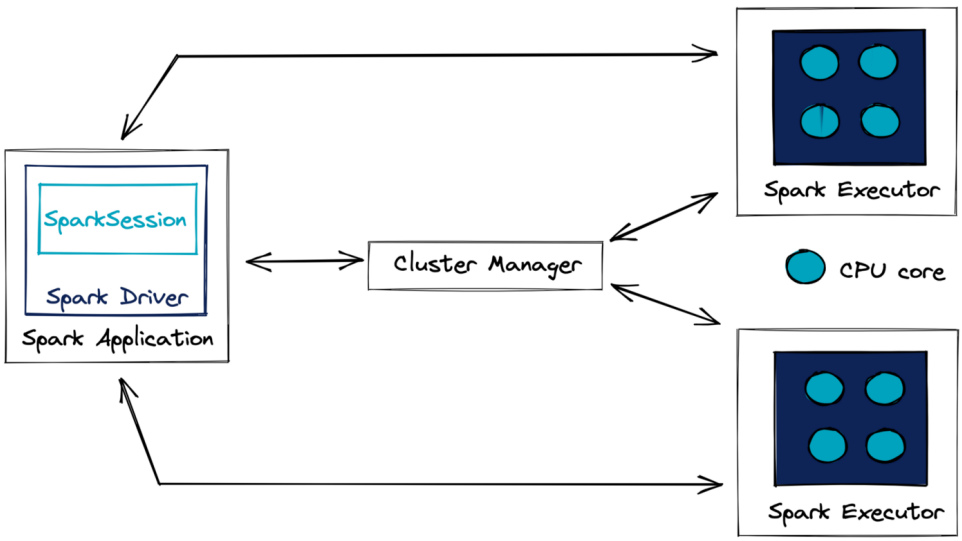


Figura 3 – Arquitetura de um Spark Cluster.
Fonte: (GRAH, 2021)

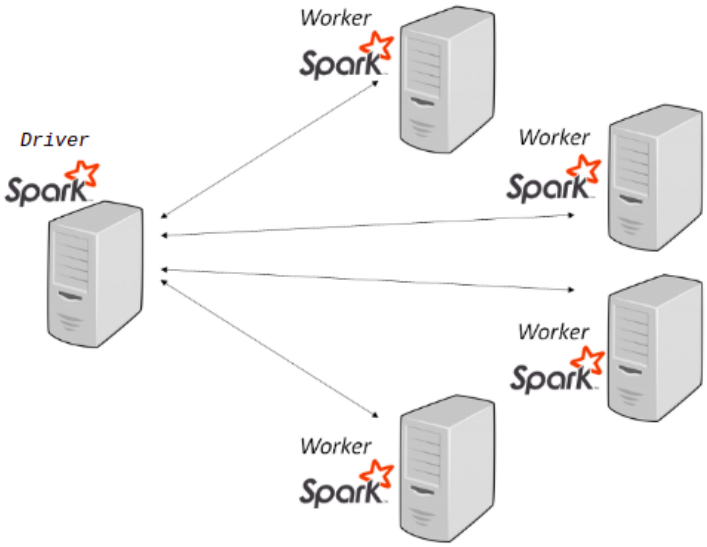


Figura 4 – Relação Spark Driver x Spark Worker.
Fonte: (KARAGOZ, 2021)

Nota-se que há duas categorias de nós em um *cluster*:

- *Driver*
- *Worker/Executor*

O *Driver* é a máquina que comanda o *cluster*, delegando tarefas, recebendo resultados e definindo quais tarefas serão feitas por quem e quando na execução de um algoritmo no *cluster*. É no *Spark Driver* em que uma sessão Spark é iniciada e onde será inicializada cada etapa da aplicação implementada. Há apenas um *Driver* no *cluster*.

Os *Workers* ou *Executors* são os responsáveis por executar as tarefas de forma paralela. Cada computador designado a ser um *worker* pode se utilizar de cada núcleo do seu processador para distribuir tarefas, ou seja, um núcleo pode ser considerado um nó onde uma tarefa será realizada. No caso deste projeto, estamos trabalhando com 20 máquinas com 4 núcleos cada, ou seja, podemos particionar e distribuir um trabalho de dados em 80 nós que executam tarefas e processam os dados ao mesmo tempo.

Por fim, há uma peça fundamental no centro do diagrama apresentado que seria o *Cluster Manager*, o ponto intermediário, que estabelece as conexões e a comunicação entre os nós. No caso da Seazone, nós utilizamos o serviço de nuvem da Amazon, o qual oferece um serviço de *cluster* denominado *Elastic MapReduce* (AWS EMR), que será descrito a seguir.

2.1.1 Outras tecnologias estudadas

Vale ressaltar que outras tecnologias capazes de fazer o processamento de dados foram estudadas, como:

- Modelo puro em Python-Pandas.
- Modelo puro em Python-Dask.
- Google Big Query.
- SynapseML com serviço em nuvem da Azure da Microsoft.

Inclusive, protótipos MVP foram realizados para cada uma delas. No entanto, o modelo puro em Pandas não foi suficiente para realizar o processamento dos dados em larga escala, sofrendo lentidão e até travamento do computador executando o algoritmo por processamento de memória excessivo. Isto porque o Pandas é uma tecnologia centralizada, executada em apenas uma máquina local, que tenta processar todos os dados de uma só vez.

O modelo puro em Dask foi ligeiramente mais promissor, já que a biblioteca Dask é capaz de particionar o conjunto de dados em subconjuntos de dados menores e realizar o processamento dos dados de forma paralela. No entanto, ainda é uma tecnologia local, executada em apenas uma máquina, por isso mesmo assim sofreu com os mesmos problemas do Pandas, lentidão e travamento.

O Google Big Query se mostrou promissor para realizar a tarefa, inclusive contava com uma interface de fácil usabilidade para o usuário e com ferramentas de modelos preditivos pré configuradas, inclusive para regressões quantílicas e logísticas. No entanto, o custo era muito elevado, inviabilizando esta tecnologia para este projeto.

Por fim, a biblioteca SynapseML também se mostrou promissora, contando com uma série de funções pré programadas para realizar tarefas de *Machine Learning*, entre elas as regressões quantílica e logística. Além disso, é uma tecnologia distribuída em nuvem, capaz de particionar e distribuir o processamento dos dados em diversas máquinas na nuvem, o que seria um diferencial frente ao Pandas e Dask, por exemplo. No entanto, por ser uma biblioteca para Python desenvolvida pela Microsoft, sua comunicação com o serviço de nuvem se dá através apenas do serviço de nuvem da Microsoft, a Azure. Como na empresa já utilizamos o serviço de nuvem da Amazon, esta característica também inviabilizou o uso desta tecnologia neste trabalho.

2.2 CLUSTER NA NUVEM - AWS EMR

Muito além de vender livros, a Amazon fornece um serviço de computadores em nuvem que conta com uma quantidade enorme de máquinas prontas para para serem utilizadas em diversas aplicações de TI de forma remota. É possível estabelecer conexão com estes computadores e executar tarefas à distância neles. É possível utilizá-los de forma individual e isolada, ou também é possível utilizar múltiplas máquinas em conjunto para realizar tarefas. Este conjunto de máquinas é denominado um *cluster*.

Neste projeto utilizamos o serviço de *cluster Elastic MapReduce* (Amazon EMR), que já possui configurações prontas específicas para realizar trabalhos utilizando PySpark. Como mencionado anteriormente, um *cluster* de máquinas em PySpark conta com uma máquina que comanda o processo (*Driver*) e as demais executam tarefas em paralelo (*Workers*), porém é necessário que haja um *cluster manager* que realize a comunicação entre elas, como apontado na figura 11 . Neste caso, é justamente o serviço da Amazon EMR que permite esta comunicação.

2.3 CLUSTER LOCAL E TECNOLOGIAS ASSOCIADAS

O *cluster* na nuvem (EMR) é um serviço pago, com custo proporcional à quantidade de máquinas utilizadas, tamanho e capacidade das máquinas (memória e processamento) e tempo de execução. Por isso, desenvolver, testar e validar os algoritmos para os modelos de categorização e predições de faturamento e ocupação utilizando este serviço seria muito custoso. Por isso, estudou-se a possibilidade de desenvolver um ambiente de trabalho em PySpark de forma local, sem utilizar o serviço de nuvem na Amazon.

Encontramos uma forma de utilizar os próprios núcleos de processadores de uma máquina local como nós de paralelismo para a distribuição de tarefas. A tecnologia, denominada de *Standalone Cluster*, foi encontrada em um repositório do desenvolvedor "mvillarrealb"e está disponível em (VILLARREAL, 2021).

Assim, foi possível utilizar um ambiente gratuito e local para desenvolver e validar MVP's dos algoritmos utilizando um subconjunto de dados pequeno, enquanto o *cluster* na EMR foi utilizado apenas quando era necessário testar e validar o algoritmo por completo, com todos os dados. Isso permitiu uma rotina de trabalho com custos minimizados de serviço de nuvem.

2.3.1 Docker

Para utilizar os núcleos dos processadores do meu computador local como nós capazes de desempenhar tarefas no ambiente em Spark é necessário configurar esses núcleos e encapsulá-los com a tecnologia do Spark.

Para realizar este encapsulamento, utilizamos a tecnologia *Docker*. O *Docker* é uma plataforma *open source* capaz de empacotar aplicações em contêineres contendo todos os componentes, bibliotecas e dependências necessárias para a execução de uma tarefa ou algoritmo dentro do próprio contêiner de forma isolada.

Assim, cada núcleo do processador (*worker*) que recebe uma tarefa a ser executada, já recebe esta tarefa em um ambiente (contêiner) completamente configurado e preparado para realizá-la.

2.4 PYTHON E BIBLIOTECAS PARA DADOS E MACHINE LEARNING

Como mencionado anteriormente, PySpark é uma API que permite utilizar a tecnologia do Apache Spark em Python. Ou seja, tudo que o Python já oferece pode ser utilizado em conjunto com o Apache Spark. Isto inclui, por exemplo, a criação e encapsulamento de funções, sintaxe comum ao Python e importação de bibliotecas para usos diversos.

No caso deste projeto, destacamos o uso de quatro bibliotecas comumente utilizadas para o trabalho com dados e *Machine Learning*:

- Sqlalchemy.
- Pandas.
- Numpy.
- Statsmodels.

Com a biblioteca *Sqlalchemy* é possível estabelecer conexão com o banco de dados da empresa para importação dos dados de *input* dos modelos.

Com a biblioteca *Pandas* é possível manipular e organizar os dados em *Dataframes*, uma estrutura que moldura os dados em linhas e colunas (similares a uma tabela) e que permite uma série de operações nos conjuntos de dados.

Com a biblioteca *Numpy* é possível realizar cálculos e operações com matrizes e vetores e também é compatível com a estrutura de dados em *Dataframes* do Pandas.

Por fim, a biblioteca *Statsmodels* conta com uma série de funções prontas para se realizar técnicas estatísticas e de álgebra linear em conjuntos de dados, as quais serão fundamentais para os modelos de categorização e predições de faturamento e ocupação dos imóveis e serão descritas com maior profundidade no capítulo 3..

2.5 GITHUB

Para o controle de versão do *software* durante a etapa de desenvolvimento foi utilizada a plataforma GitHub. Nela é possível hospedar todos os códigos e arquivos necessários para o projeto em um repositório aberto em que outros integrantes da equipe de dados possam acessá-los para revisá-los e também contribuir.

2.6 DBEAVER

Para a visualização das tabelas geradas pelos modelos e armazenadas no banco de dados da empresa utilizamos o DBeaver, um software cliente SQL capaz de ler, escrever e editar dados em um banco de dados relacional.

2.7 POWER BI

Por fim, para a disponibilização dos resultados finais dos modelos em *dashboards* contendo tabelas, gráficos e indicadores utilizou-se o Power BI. Foram desenvolvidos *dashboards* de validação dos resultados dos modelos e *dashboards* que resumem e disponibilizam os resultados para que possam ser de fato utilizados por diferentes setores da empresa no dia a dia.

3 DESCRIÇÃO DOS ALGORITMOS

Neste capítulo são apresentadas as arquiteturas e lógicas dos algoritmos implementados para os modelos de categorização dos imóveis e para os modelos de predição de faturamento e taxa de ocupação dos imóveis do Airbnb, bem como etapas intermediárias e detalhes fundamentais para suas implementações.

3.1 DEFINIÇÃO DE CENÁRIO

Tanto para o modelo de categorização de imóveis quanto para os modelos de predição de faturamento e taxa de ocupação foi necessário segregar os resultados dos modelos por algumas circunstâncias que fossem coerentes. Para a clusterização de preço, há inúmeras circunstâncias que podem fazer o preço da diária de um imóvel variar, tratando-se de um mesmo imóvel individualmente ou de um conjunto de imóveis. O mesmo se aplica para a ocupação de um imóvel. O estudo dessas dinâmicas é fundamental para que sejam levadas em consideração na construção dos modelos.

Por isso, um conjunto de circunstâncias foi levantado e segregado. A este conjunto, denominou-se de *cenário*. Ao estudarmos o que poderia fazer o preço da diária e a ocupação de um imóvel variar, levantou-se:

- Localização do imóvel.
- Tipo do imóvel (casa, apartamento, hotel ou outros).
- Número de quartos do imóvel.
- Sazonalidade da diária.

Esta segregação se mostrou coerente e cada ponto será aprofundado e exemplificado a seguir.

Para o caso da localização do imóvel, temos que cada localidade possui preços de diárias diferentes. Bairros com maior demanda no Airbnb tendem a possuir valores de diárias maiores do que bairros com menor demanda, por exemplo, além de também receberem um volume de reservas maior, afetando a taxa de ocupação desses imóveis. A demanda de um imóvel em uma localização pode variar de acordo com diversos fatores, como pontos turísticos no entorno do imóvel, taxas de violência e assaltos no bairro do imóvel, fácil acesso a meios de transporte, restaurantes, farmácias, bancos e outras comodidades, entre outros fatores. Por isso, levar a localização do imóvel em consideração na definição de cenário mostrou-se fundamental.

Também é fundamental considerar o tipo do imóvel na definição de cenário, já que casas, apartamentos e hotéis tendem a ser precificados separadamente, geralmente comparados entre seus concorrentes do mesmo tipo. Além disso, a demanda

por reservas também varia para estes diferentes tipos de imóvel no Airbnb, afetando a taxa de ocupação destes grupos de imóveis.

O número de quartos do imóvel é inserido na definição de cenário, já que existe uma tendência linear entre o número de quartos e o preço da diária, como pode ser visto na imagem a seguir:

Figura 5 – Número de quartos do imóvel x Preço médio da diária.
Fonte: Medium

A demanda de reserva também varia nesses casos, por se tratar de uma configuração diferente de imóvel e que afeta diretamente a configuração do público que será hospedado.

Por fim, a sazonalidade entra na definição de cenário já que a demanda por imóveis também varia ao longo do ano. Em regiões praianas, por exemplo, a demanda por imóveis no Airbnb tende a ser maior em meses quentes do ano, enquanto que regiões em áreas de montanha no interior tendem a ter uma demanda maior nos meses mais frios. Essa dinâmica afeta diretamente os preços e as reservas dos imóveis.

Esta definição de cenário afetará diretamente a escolha das variáveis de entrada do modelo, o que será descrito mais profundamente a seguir.

3.2 DEFINIÇÃO DAS VARIÁVEIS DE ENTRADA

A própria definição de cenário mencionada anteriormente já sugere quais variáveis serão utilizadas como entrada para os modelos.

Por isso, temos que as variáveis de entrada para o modelo de regressão quantílica (o qual será utilizado tanto para a categorização dos imóveis quanto para a

predição de faturamento) definidas são:

Variável	**Tipo do dado**
Cidade	String
Bairro	String
Tipo do imóvel	String
Número de quartos	Int
Data	Datetime
Mês	String
Feriado	Booleano
Fim de semana	Booleano
Preço	Float

Tabela 1 – Variáveis de entrada do modelo de Regressão Quantílica.

Sendo que preço é a variável da regressão.

É importante ressaltar que o resultado do modelo de categorização dos imóveis, ou seja, a categoria definida para cada imóvel será utilizada como variável de entrada para o modelo de predição de taxa de ocupação dos imóveis, já que queremos a ocupação separada por categoria. Por isso, as variáveis de entrada definidas para o modelo de regressão logística são:

Variável	**Tipo do dado**
Cidade	String
Bairro	String
Tipo do imóvel	String
Número de quartos	Int
Data	Datetime
Mês	String
Feriado	Booleano
Fim de semana	Booleano
Diária ocupada	Booleano
Categoria do imóvel	String

Tabela 2 – Variáveis de entrada do modelo de Regressão Logística.

Já para o modelo de Regressão Logística a variável da regressão é a Diária ocupada.

3.3 BASE DE DADOS E PROCEDIMENTOS DE LIMPEZA E TRATAMENTO DE DADOS

A base de dados utilizada para a construção dos modelos conta com quatro tabelas armazenadas no banco de dados da empresa:

Tabela	**Descrição**
Details	Informações sobre os listings como número de quartos e tipo do imóvel.
Location	Localização dos listings, com foco em cidade e bairro.
Block and occupancy	Informações sobre as reservas como preço da diária, diária reservada e datas bloqueadas.
Dates	Informações sobre as datas como feriados e fins de semana.

Tabela 3 – Descrição das tabelas da base de dados utilizadas.

Os procedimentos de limpeza e tratamento implementados nos dados de entrada dos modelos foram os seguintes:

- Exclusão de *listings* sem bairro.
- Exclusão de *listings* com avaliações abaixo de 4 estrelas no Airbnb.
- Agrupamento de bairros pequenos de Florianópolis em um conjunto de bairros. Por exemplo: Pantanal, Trindade e Carvoeira tornam-se bairro UFSC.
- Agrupamento de casas e hotéis com mais de oito quartos.
- Agrupamento de apartamentos com mais de cinco quartos.
- Feriado em *string* para feriado em booleano.
- Mês em inteiro para mês em *string*. Este tratamento é de extrema importância, pois afeta diretamente o cálculo dos modelos. Será explicado mais a fundo posteriormente.
- Duplicação dos dados atribuindo o valor de bairro "ALL"para os *listings* para calcular um modelo por bairro e um por cidade.
- Exclusão de registros com campos nulos/vazios.

3.4 MODELO DE CATEGORIZAÇÃO DE IMÓVEIS DO AIRBNB

Como mencionado anteriormente, o modelo de categorização de imóveis do Airbnb tem como objetivo receber um conjunto de dados contendo os imóveis do Airbnb e atribuir uma categoria para cada um dos imóveis, como definido na seção 1.2 (Categorias: SIM, JR, SUP, TOP, MASTER).

Para isso, será aplicado um modelo de regressão quantílica no preço da diária dos imóveis, levando em consideração as variáveis de entrada mencionadas anteriormente. Este método é capaz de definir quantis de preço, ou seja, definir limites de preço que dividam o intervalo de frequência da amostra de dados das diárias dos imóveis em partes bem definidas, o que resultará em uma divisão de grupos de imóveis em faixas

de preço, separados em diversos cenários (como definido em 3.1). Os imóveis que forem mapeados em cada faixa serão atribuídos à categoria respectiva àquela faixa de preço.

Vale ressaltar que a regressão quantílica será aplicada no logaritmo do preço (log(preço)). Isto porque a distribuição do preço das diárias dos imóveis no Airbnb é uma distribuição logarítmica normal, como evidenciado na figura 6:

Figura 6 – Distribuição logarítmica normal dos preços das diárias dos imóveis no Airbnb. Fonte: airbnb.com.br

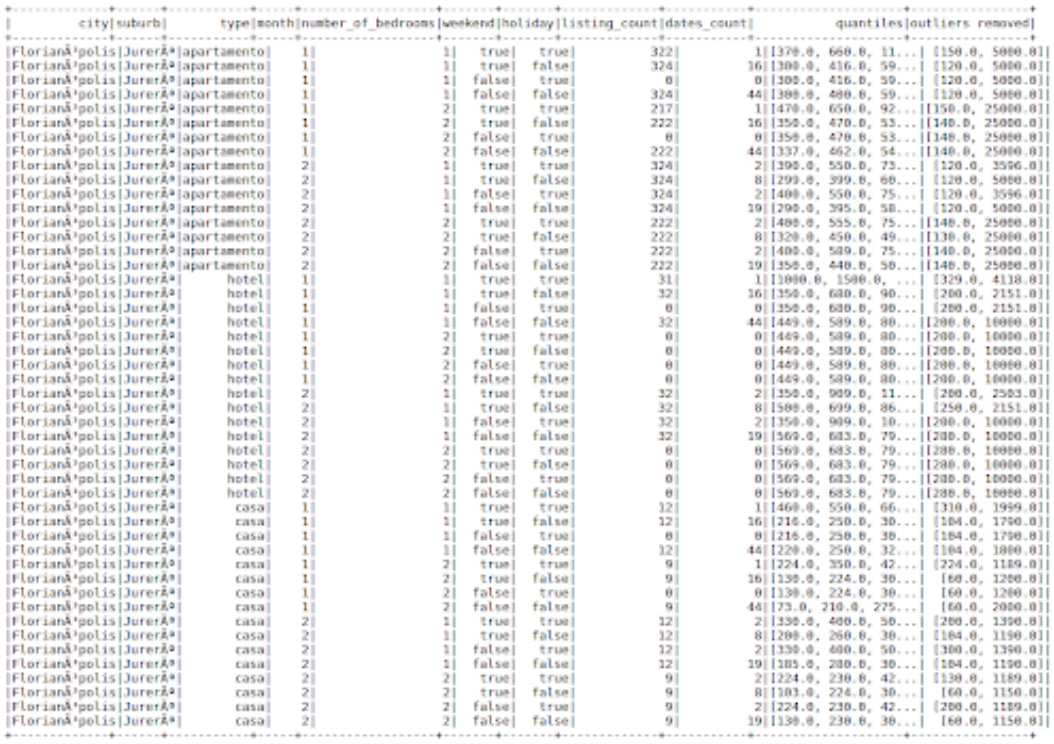
3.4.1 Quantis definidos para o Modelo de Regressão Quantílica

Os quantis escolhidos para a definição das faixas de preço de cada categoria foram:

- Quantil de 0,1% a 20%: Categoria SIM (Simples).
- Quantil de 20% a 40%: Categoria JR (Júnior).
- Quantil de 40% a 60%: Categoria SUP (Superior).
- Quantil de 60% a 80%: Categoria TOP (Top).
- Quantil de 80% a 99,9%: Categoria MASTER (Master).

Assim, ao aplicar o modelo de regressão quantílica no logaritmo do preço, para estes quantis definidos, obtém-se cinco faixas de preço para cada cenário exclusivo presente no banco de dados.

Esta etapa está retratada na figura 7, onde é possível ver os resultados divididos por cenário (cidade, bairro, mês, número de quartos, tipo do imóvel, tipo da diária), como descrito anteriormente.



city	suburb	type	month	number_of_bedrooms	weekend	holiday	listing_count	dates_count	quantiles	outliers removed
Florianã³polis	Jurerã*	apartamento	1	1	true	true	322	1	[370.0, 660.0, 11...	[150.0, 5000.0]
Florianã³polis	Jurerã³	apartamento	1	1	true	false	324	16	[300.0, 416.0, 59...	[120.0, 5000.0]
Florianã³polis	Jurerã*	apartamento	1	1	false	true	0	0	[300.0, 416.0, 59...	[120.0, 5000.0]
Florianã³polis	Jurerã*	apartamento	1	1	false	false	324	44	[300.0, 400.0, 59...	[120.0, 5000.0]
Florianã³polis	Jurerã³	apartamento	1	2	true	true	217	1	[470.0, 650.0, 92...	[150.0, 25000.0]
Florianã³polis	Jurerã³	apartamento	1	2	true	false	222	16	[350.0, 470.0, 53...	[140.0, 25000.0]
Florianã³polis	Jurerã*	apartamento	1	2	false	true	0	0	[350.0, 470.0, 53...	[140.0, 25000.0]
Florianã³polis	Jurerã³	apartamento	1	2	false	false	222	44	[337.0, 462.0, 54...	[140.0, 25000.0]
Florianã³polis	Jurerã³	apartamento	2	1	true	true	324	2	[390.0, 550.0, 73...	[120.0, 3596.0]
Florianã³polis	Jurerã*	apartamento	2	1	true	false	324	8	[293.0, 399.0, 60...	[120.0, 5000.0]
Florianã³polis	Jurerã*	apartamento	2	1	false	true	324	2	[400.0, 550.0, 75...	[120.0, 3596.0]
Florianã³polis	Jurerã³	apartamento	2	1	false	false	324	19	[290.0, 395.0, 58...	[120.0, 5000.0]
Florianã³polis	Jurerã*	apartamento	2	2	true	true	222	2	[400.0, 555.0, 75...	[140.0, 25000.0]
Florianã³polis	Jurerã³	apartamento	2	2	true	false	222	8	[320.0, 450.0, 49...	[130.0, 25000.0]
Florianã³polis	Jurerã³	apartamento	2	2	false	true	222	2	[400.0, 589.0, 75...	[140.0, 25000.0]
Florianã³polis	Jurerã*	apartamento	2	2	false	false	222	19	[350.0, 440.0, 58...	[140.0, 25000.0]
Florianã³polis	Jurerã*	hotel	1	1	true	true	31	1	[1000.0, 1500.0, ...	[329.0, 4138.0]
Florianã³polis	Jurerã³	hotel	1	1	true	false	32	16	[350.0, 600.0, 90...	[200.0, 2151.0]
Florianã³polis	Jurerã*	hotel	1	1	false	true	0	0	[350.0, 600.0, 90...	[200.0, 2151.0]
Florianã³polis	Jurerã*	hotel	1	1	false	false	32	44	[449.0, 589.0, 80...	[200.0, 10000.0]
Florianã³polis	Jurerã³	hotel	1	2	true	true	0	0	[449.0, 589.0, 80...	[200.0, 10000.0]
Florianã³polis	Jurerã³	hotel	1	2	true	false	0	0	[449.0, 589.0, 80...	[200.0, 10000.0]
Florianã³polis	Jurerã*	hotel	1	2	false	true	0	0	[449.0, 589.0, 80...	[200.0, 10000.0]
Florianã³polis	Jurerã³	hotel	1	2	false	false	0	0	[449.0, 589.0, 80...	[200.0, 10000.0]
Florianã³polis	Jurerã³	hotel	2	1	true	true	32	2	[350.0, 909.0, 11...	[200.0, 2503.0]
Florianã³polis	Jurerã*	hotel	2	1	true	false	32	8	[500.0, 699.0, 86...	[250.0, 2151.0]
Florianã³polis	Jurerã*	hotel	2	1	false	true	32	2	[350.0, 909.0, 10...	[200.0, 10000.0]
Florianã³polis	Jurerã³	hotel	2	1	false	false	32	19	[569.0, 683.0, 79...	[200.0, 10000.0]
Florianã³polis	Jurerã³	hotel	2	2	true	true	0	0	[569.0, 683.0, 79...	[200.0, 10000.0]
Florianã³polis	Jurerã³	hotel	2	2	true	false	0	0	[569.0, 683.0, 79...	[200.0, 10000.0]
Florianã³polis	Jurerã³	hotel	2	2	false	true	0	0	[569.0, 683.0, 79...	[200.0, 10000.0]
Florianã³polis	Jurerã*	hotel	2	2	false	false	0	0	[569.0, 683.0, 79...	[200.0, 10000.0]
Florianã³polis	Jurerã³	casa	1	1	true	true	12	1	[460.0, 550.0, 66...	[310.0, 1999.0]
Florianã³polis	Jurerã³	casa	1	1	true	false	12	16	[216.0, 250.0, 30...	[104.0, 1750.0]
Florianã³polis	Jurerã*	casa	1	1	false	true	0	0	[216.0, 250.0, 30...	[104.0, 1750.0]
Florianã³polis	Jurerã³	casa	1	1	false	false	12	44	[220.0, 250.0, 32...	[104.0, 1800.0]
Florianã³polis	Jurerã³	casa	1	2	true	true	9	1	[224.0, 350.0, 42...	[224.0, 1189.0]
Florianã³polis	Jurerã*	casa	1	2	true	false	9	16	[130.0, 224.0, 30...	[60.0, 1200.0]
Florianã³polis	Jurerã³	casa	1	2	false	true	0	0	[130.0, 224.0, 30...	[60.0, 1200.0]
Florianã³polis	Jurerã³	casa	1	2	false	false	9	44	[73.0, 210.0, 275...	[60.0, 2000.0]
Florianã³polis	Jurerã*	casa	2	1	true	true	12	2	[330.0, 400.0, 50...	[200.0, 1390.0]
Florianã³polis	Jurerã*	casa	2	1	true	false	12	8	[200.0, 260.0, 30...	[104.0, 1190.0]
Florianã³polis	Jurerã³	casa	2	1	false	true	12	2	[330.0, 400.0, 50...	[300.0, 1390.0]
Florianã³polis	Jurerã*	casa	2	1	false	false	12	19	[185.0, 280.0, 30...	[104.0, 1190.0]
Florianã³polis	Jurerã³	casa	2	2	true	true	9	2	[224.0, 230.0, 42...	[130.0, 1189.0]
Florianã³polis	Jurerã³	casa	2	2	true	false	9	8	[103.0, 224.0, 30...	[60.0, 1150.0]
Florianã³polis	Jurerã*	casa	2	2	false	true	9	2	[224.0, 230.0, 42...	[200.0, 1189.0]
Florianã³polis	Jurerã*	casa	2	2	false	false	9	19	[130.0, 230.0, 30...	[60.0, 1150.0]

Figura 7 – Quantis de preço obtidos pelo modelo para cada cenário.

Vale ressaltar também, que o próprio modelo de regressão quantílica permite a remoção de *outliers* do modelo, os quais estão explícitos na última coluna da figura 7. Neste caso, foram definidas como *outliers* e desconsideradas do cálculo das faixas de preço as diárias que se encontraram nos quantis 0,1% e 99,9%.

3.4.2 *Score*/Pontuação das diárias e definição da categoria final

Desta forma, temos que todas as diárias dos imóveis são mapeadas dentro destas cinco faixas e a cada diária atribui-se uma pontuação:

- Faixa SIM: 1 ponto.
- Faixa JR: 2 pontos.
- Faixa SUP: 3 pontos.
- Faixa TOP: 4 pontos.
- Faixa MASTER: 5 pontos.

Assim, para obter a categoria final de um *listing*, calcula-se a média da pontuação das diárias (arredondada) de cada *listing*, atingindo um valor médio final para o *listing*. Desta maneira, define-se a categoria final como:

- 0 <= Pontuação < 1,5: Listing SIM (Simples).

- 1,5 <= Pontuação < 2,5: Listing JR (Júnior).
- 0 <= Pontuação < 3,5: Listing SUP (Superior).
- 0 <= Pontuação < 4,5: Listing TOP (Top).
- 0 <= Pontuação <= 5: Listing MASTER (Master).

Esta etapa está retratada na figura 8, onde é possível ver, para cada *listing*, quantas diárias foram alocadas em cada faixa de preço calculada pelo modelo, o *score* final de cada *listing* e a categoria atribuída.

airbnb_listing_id	SIM_count	JR_count	SUP_count	TOP_count	MASTER_count	score	strata_out	Run date
47990968	0	35	10	37	12	3.0	SUP	2022-02-18 18:51:...
717432	0	0	0	0	54	5.0	MASTER	2022-02-18 18:51:...
30149793	37	57	0	0	0	2.0	JR	2022-02-18 18:51:...
44534631	33	2	38	18	3	3.0	SUP	2022-02-18 18:51:...
15446776	57	37	0	0	0	1.0	SIM	2022-02-18 18:51:...
21935899	0	81	13	0	0	2.0	JR	2022-02-18 18:51:...
38698647	0	0	1	4	54	5.0	MASTER	2022-02-18 18:51:...
26448101	64	14	14	2	0	2.0	JR	2022-02-18 18:51:...
8183766	10	25	1	1	57	4.0	TOP	2022-02-18 18:51:...
21714868	6	50	0	5	0	2.0	JR	2022-02-18 18:51:...
25454572	0	66	27	1	0	2.0	JR	2022-02-18 18:51:...
16527384	0	0	0	1	93	5.0	MASTER	2022-02-18 18:51:...
4543873	0	35	5	53	1	3.0	SUP	2022-02-18 18:51:...
29832903	0	0	0	1	93	5.0	MASTER	2022-02-18 18:51:...
23888107	0	0	0	38	56	5.0	MASTER	2022-02-18 18:51:...
40582204	0	0	5	75	14	4.0	TOP	2022-02-18 18:51:...
38784359	0	0	1	28	65	5.0	MASTER	2022-02-18 18:51:...
42543605	0	0	0	49	45	4.0	TOP	2022-02-18 18:51:...
28569932	45	43	6	0	0	2.0	JR	2022-02-18 18:51:...
49910878	0	35	8	13	38	4.0	TOP	2022-02-18 18:51:...
44664571	4	31	1	4	54	4.0	TOP	2022-02-18 18:51:...
32612246	27	12	7	7	41	3.0	SUP	2022-02-18 18:51:...
40593983	5	32	22	0	0	2.0	JR	2022-02-18 18:51:...
4820995	0	0	0	4	90	5.0	MASTER	2022-02-18 18:51:...
24986690	0	1	78	15	0	3.0	SUP	2022-02-18 18:51:...
39721155	32	60	2	0	0	2.0	JR	2022-02-18 18:51:...
24986856	0	1	83	18	0	3.0	SUP	2022-02-18 18:51:...
42814765	91	1	0	2	0	1.0	SIM	2022-02-18 18:51:...
39642127	11	83	0	0	0	2.0	JR	2022-02-18 18:51:...
16954539	0	0	0	0	94	5.0	MASTER	2022-02-18 18:51:...
14336569	11	12	32	9	7	3.0	SUP	2022-02-18 18:51:...
23686828	54	40	0	0	0	1.0	SIM	2022-02-18 18:51:...
30232277	40	35	15	1	3	2.0	JR	2022-02-18 18:51:...
2159840	67	23	1	3	0	1.0	SIM	2022-02-18 18:51:...
2061157	0	5	28	9	52	4.0	TOP	2022-02-18 18:51:...
31414733	89	5	0	0	0	1.0	SIM	2022-02-18 18:51:...
4828196	39	18	12	24	1	2.0	JR	2022-02-18 18:51:...
31191932	0	0	5	0	89	5.0	MASTER	2022-02-18 18:51:...
42362406	0	0	0	0	35	5.0	MASTER	2022-02-18 18:51:...
27886998	0	33	1	51	9	3.0	SUP	2022-02-18 18:51:...
17148942	1	73	20	0	0	2.0	JR	2022-02-18 18:51:...
46374895	0	0	0	49	45	4.0	TOP	2022-02-18 18:51:...
52512375	0	0	4	76	14	4.0	TOP	2022-02-18 18:51:...
40701296	0	0	5	32	57	5.0	MASTER	2022-02-18 18:51:...
51684548	0	0	0	6	88	5.0	MASTER	2022-02-18 18:51:...
46118964	0	1	15	27	51	4.0	TOP	2022-02-18 18:51:...
44289992	0	4	37	0	53	4.0	TOP	2022-02-18 18:51:...
49985665	5	0	89	0	0	3.0	SUP	2022-02-18 18:51:...
46835662	0	0	1	6	87	5.0	MASTER	2022-02-18 18:51:...
58027972	0	0	4	76	14	4.0	TOP	2022-02-18 18:51:...
14273778	43	26	23	2	0	2.0	JR	2022-02-18 18:51:...

Figura 8 – Categoria final dos *listings* definidas por cálculo da pontuação das diárias.

3.4.3 Fluxograma - Modelo de Categorização

Para fins de documentação do algoritmo, desenvolveu-se o fluxograma da figura 9 utilizando a aplicação web *Lucidchart* para o modelo de categorização:

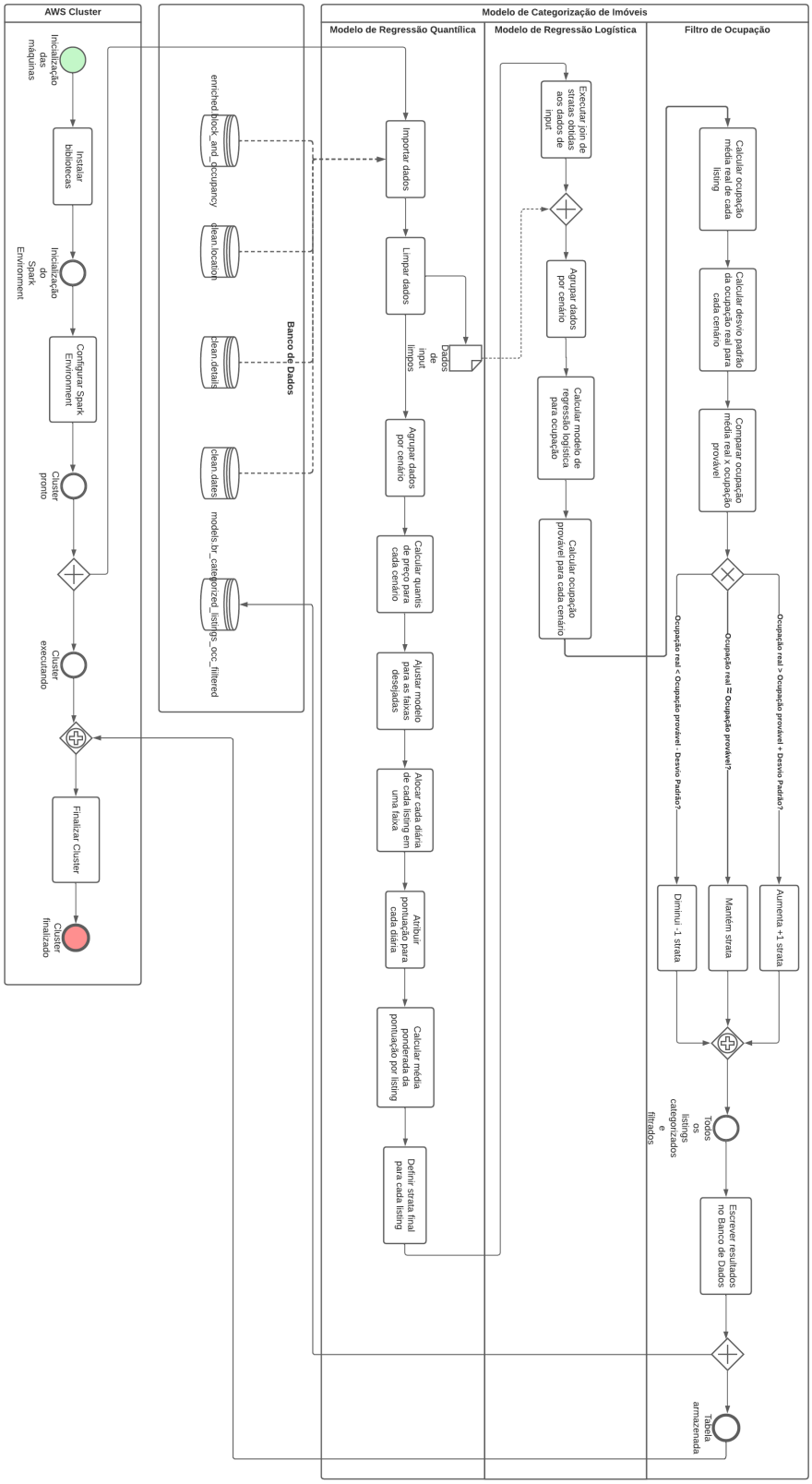


Figura 9 – Fluxograma do modelo de categorização de imóveis.

O fluxograma foi desenvolvido seguindo as normas BPMN 2.0 apresentadas em (SERVIÇO PÚBLICO DO ESPÍRITO SANTO, 2019) e representa as principais etapas do algoritmo:

1. Importação dos dados das 4 tabelas de entrada do banco de dados (descritas na tabela 3).
2. Limpezas dos dados (descritas na seção 3.3).
3. Agrupamento dos dados por cenário (descrito na seção 3.1).
4. Cálculo dos quantis para cada cenário e obtenção das faixas de preço (descrito na seção 3.4.1).
5. Alocação das diárias em suas respectivas faixas.
6. Pontuação das diárias de acordo com a faixa alocada (descrito na seção 3.4.2).
7. Cálculo do *score* de cada *listing* com base na pontuação de suas diárias (também descrito na seção 3.4.2).
8. Definição da categoria do *listing*.
9. Importação do resultado do modelo de predição de taxa de ocupação.
10. Aplicação do filtro de taxa de ocupação para corrigir categoria de *listings* mal precificados (descrito na seção 3.6).
11. Definição da categoria final de cada *listing*.
12. Registro dos resultados no banco de dados.

As etapas 9 e 10 serão descritas com maior profundidade a seguir, na seção 3.6.

3.5 MODELOS DE PREDIÇÃO DE FATURAMENTO E TAXA DE OCUPAÇÃO DE IMÓVEIS DO AIRBNB

Como mencionado anteriormente, o modelo de predição de taxa de ocupação dos imóveis do Airbnb tem como objetivo receber um conjunto de dados contendo os imóveis do Airbnb, as diárias de cada imóvel e suas respectivas ocupações (diária ocupada ou livre) e, a partir destas informações, agrupar e segregar este conjunto de dados em cenários e fornecer uma predição de taxa de ocupação de cada cenário. Isto é, a predição da taxa de ocupação não é realizada imóvel a imóvel, mas sim a um grupo de imóveis similares separados por cidade, bairro, tipo do imóvel, número de quartos e categoria do imóvel e também ao longo do tempo, fornecendo predições para cada

mês do ano para estes grupos. Inclusive, por este motivo primeiro é calculado o modelo de categorização dos imóveis utilizando a regressão quantílica e a categoria obtida para cada imóvel é utilizada como variável de entrada para o modelo de regressão logística para realizar a predição da taxa de ocupação dos imóveis de cada cenário exclusivo presente no banco de dados.

Já para o modelo de predição de faturamento, os modelos de regressão quantílica e de regressão logística são combinados para realizar o cálculo de faturamento por cenário. Como no modelo de regressão quantílica são calculadas faixas de preço com base nos quantis de 0,1%, 20%, 40%, 60%, 80% e 99,9% de cada cenário e no modelo de regressão logística são calculadas as taxas de ocupação prováveis para cada cenário, é possível calcular quantis intermediários de preço para cada faixa, obtendo quantis médios de preço para cada categoria dos imóveis (SIM, JR, SUP, TOP, MASTER), separado por cenário, e combiná-los à taxa de ocupação provável para cada cenário, obtendo uma predição de faturamento por cenário. Como cada cenário é calculado mês a mês, temos então o cálculo para a predição de faturamento explícito na equação:

$$P_{fat_i} = P_{oc_i} * Q_{mdio_i} * d_i \tag{1}$$

Em que i representa cada cenário, P_{fat_i} é a predição de faturamento de um cenário, P_{oc_i} é a predição de taxa de ocupação de um cenário, d_i é o número de dias no mês respectivo ao cenário e Q_{mdio_i} é o quantil médio de preço do cenário. Vale ressaltar que para estes modelos a categoria dos imóveis já é considerada na composição do cenário, sendo que os quantis médios de cada categoria são dados por:

- Quantil 10% de preço: quantil médio de preço de diária para imóveis SIM (Simples) de cada cenário.
- Quantil 30% de preço: quantil médio de preço de diária para imóveis JR (Júnior) de cada cenário.
- Quantil 50% de preço: quantil médio de preço de diária para imóveis SUP (Superior) de cada cenário.
- Quantil 70% de preço: quantil médio de preço de diária para imóveis TOP (Top) de cada cenário.
- Quantil 90% de preço: quantil médio de preço de diária para imóveis MASTER (Master) de cada cenário.

3.5.1 Fluxograma - Modelos de Predição de Faturamento e Taxa de Ocupação de Imóveis do Airbnb

Para fins de documentação do algoritmo, desenvolveu-se o fluxograma da figura 10 utilizando a aplicação web *Lucidchart* para os modelo de predição de faturamento e taxa de ocupação dos imóveis:

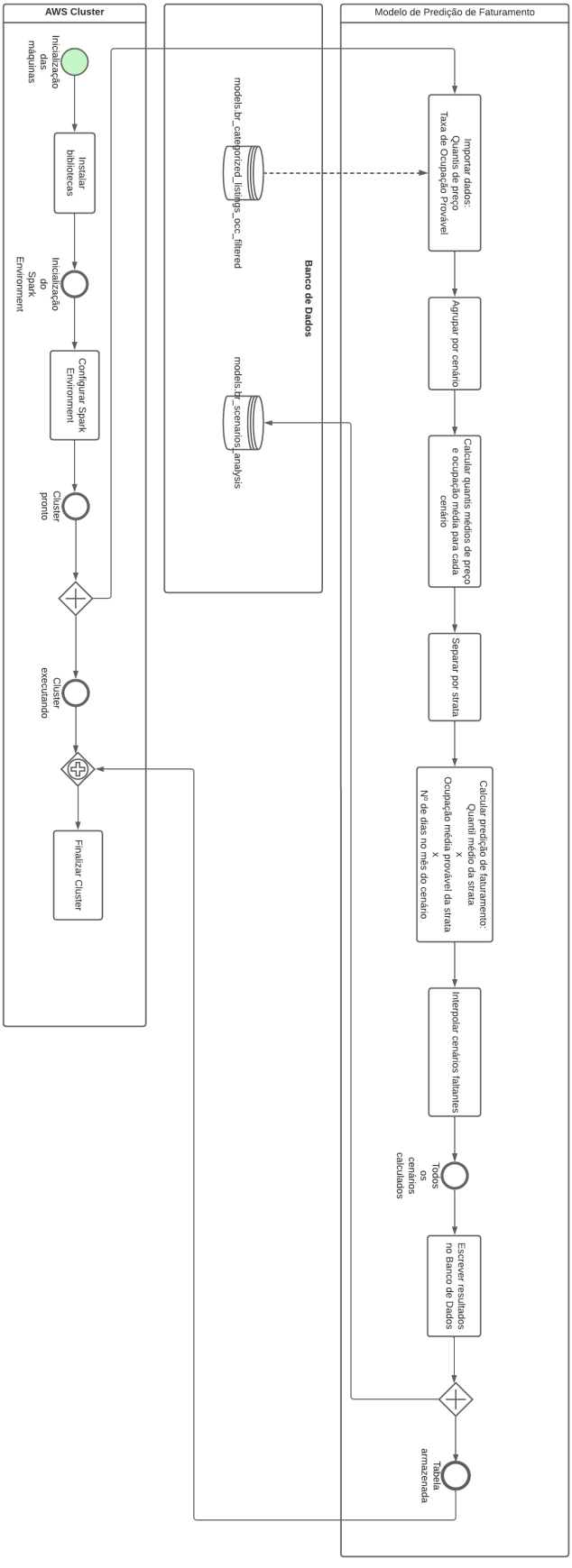


Figura 10 – Fluxograma dos modelos de predição de faturamento e taxa de ocupação dos imóveis.

O fluxograma foi desenvolvido seguindo as normas BPMN 2.0 e representa as principais etapas do algoritmo:

1. Importar os quantis médios resultantes (referenciados em 3.5) calculados no modelo de regressão quantílica.
2. Importar taxas de ocupações prováveis calculadas no modelo de regressão logística.
3. Agrupar os dados por cenário.
4. Separar os dados por categoria (definidas no modelo de categorização).
5. Aplicar o cálculo referenciado em 1.

Ao final, quando o cálculo aplicado para todos os cenários, obtém-se a predição de faturamento de cada cenário para cada categoria de imóvel.

3.6 FILTRO DE TAXA DE OCUPAÇÃO APLICADO AO MODELO DE CATEGORIZAÇÃO

Não apenas o modelo de categorização dos imóveis possui utilidade para o modelo de predição de taxa de ocupação (fornecendo a categoria como variável de entrada), como o modelo de predição de taxa de ocupação também possui utilidade para o modelo de categorização, possibilitando um ajuste fino ao final do algoritmo, o qual será descrito mais profundamente a seguir.

Apesar deste trabalho partir da premissa de mercado eficiente, como descrito na seção 1.4.1, é necessário considerar que existe a possibilidade de que em alguns casos as pessoas não realizem adequadamente a precificação dos seus imóveis - atribuindo um valor de preço de diária desproporcional à qualidade do imóvel, tanto para cima quanto para baixo.

Como a categoria do imóvel será definida a partir do modelo de regressão quantílica aplicado ao preço, este fator pode influenciar diretamente na atribuição de uma categoria inadequada para imóveis mal precificados.

Por este motivo, implementou-se um filtro de taxa de ocupação que busca corrigir eventuais problemas decorridos desta possibilidade. A hipótese é de que imóveis com preço das diárias muito acima do esperado, tendem a ter uma taxa de ocupação também muito menor do que o esperado a imóveis da mesma categoria e do mesmo cenário. Analogamente, imóveis com preço das diárias muito abaixo do esperado, tendem a ter uma taxa de ocupação maior do que o esperado a imóveis da mesma categoria e do mesmo cenário.

Por isso, inicialmente o modelo é calculado, define-se categorias preliminares a todos os imóveis e, por fim, aplica-se o filtro de taxa de ocupação podendo subir ou descer uma categoria de um imóvel.

O cálculo por trás deste filtro é feito comparando a taxa de ocupação real dos imóveis, o desvio padrão da taxa de ocupação real dos imóveis de cada cenário e categoria, com a taxa de ocupação provável calculada pelo modelo de predição de taxa de ocupação. Assim, temos que:

$$Oc_{real} > Oc_{modelo} + \alpha\sigma_{real} \rightarrow Aumenta + 1 Categoria \tag{2}$$

$$Oc_{real} < Oc_{modelo} - \alpha\sigma_{real} \rightarrow Diminui - 1 Categoria \tag{3}$$

Sendo que para determinar α alguns testes foram realizados e definiu-se que α = 1 seria ideal, assim teríamos que, com um desvio padrão, aproximadamente 68% dos *listings* teriam sua categoria preservada, aproximadamente 16% dos *listings* seriam promovidos de categoria, por estarem performando acima da média de ocupação dos *listings* de seu respectivo cenário e aproximadamente 16% dos *listings* seriam rebaixados uma categoria por estarem performando abaixo da média de ocupação dos *listings* de seu respectivo cenário.

4 RESULTADOS

Neste capítulo são apresentados os resultados dos modelos de categorização de imóveis e dos modelos de predição de faturamento e taxa de ocupação dos imóveis do Airbnb. Vale relembrar que a metodologia utilizada para os resultados parciais de validação é a mencionada na seção 1.4.3 sobre MVPs. Para resultados oficiais, o resultado completo é apresentado.

4.1 MODELO DE CATEGORIZAÇÃO DE IMÓVEIS

Inicialmente, são apresentados os resultados e o processo de validação do modelo de categorização de imóveis utilizando a regressão quantílica e o filtro da taxa de ocupação. Vale ressaltar que o modelo de predição de taxa de ocupação é utilizado na implementação do filtro, ou seja, há um cruzamento dos modelos neste ponto.

4.1.1 1° MVP - Validação da Regressão Quantílica

Inicialmente, para realizar um teste preliminar com o objetivo de validar o modelo de regressão quantílica como suficiente para separar os imóveis em faixas de preço e, por consequência, em categorias, pensou-se em calcular o modelo para apenas um cenário específico antes de expandir o modelo para todos os dados e cenários existentes no banco de dados.

Assim, o cenário definido para este MVP, seguindo as diretrizes de cenário definidas em 3.1, as variáveis de entrada definidas em 3.2 e os quantis definidos em 3.4.1, temos o resultado da regressão quantílica aplicados em imóveis do tipo 'apartamento', contendo um quarto, em Florianópolis, no mês de Janeiro:

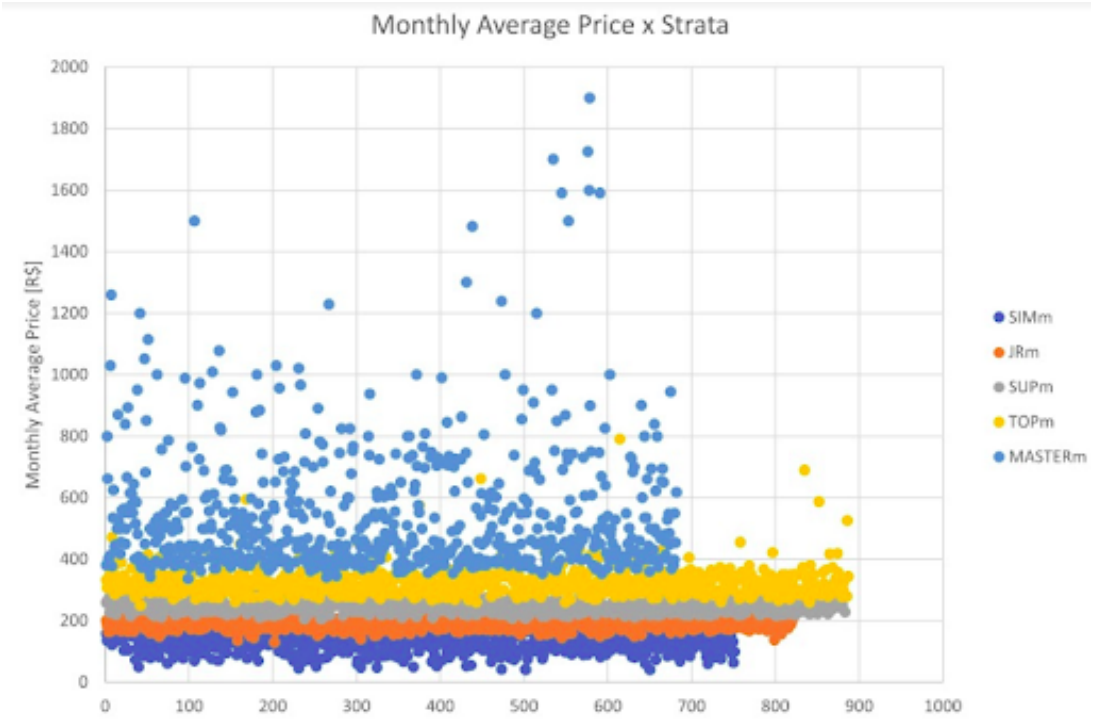


Figura 11 – Resultado do Modelo de Categorização - Faixas de preço por categoria.

Neste gráfico, cada ponto representa um imóvel. No eixo y temos o valor mensal médio da diária de um imóvel e no eixo x temos o número de imóveis para cada categoria, ou seja, nota-se de imediato que a distribuição de categorias é próxima a uma distribuição normal, com as categorias de ponta (SIM e MASTER) com menor quantidade de imóveis e a maior concentração de imóveis nas categorias intermediárias. Nota-se também que aplicar um modelo de regressão quantílica foi suficiente para separar os imóveis em faixas de preço e atribuir uma categoria para grupos de imóveis. Além disso, as proporções esperadas foram mantidas, ou seja, os imóveis de menor preço foram alocados nas categorias mais baixas, imóveis intermediários em categorias intermediárias e os imóveis mais caros nas categorias mais altas.

4.1.2 2° MVP - Validação do Filtro de Taxa de Ocupação

Após o primeiro resultado do modelo de categorização, nota-se que a regressão quantílica apresenta resultados satisfatórios para a alocação de imóveis em faixas de preço, no entanto, este não seria o resultado final das categorias. Após a alocação dos imóveis em suas categorias preliminares, o filtro de taxa de ocupação é aplicado para corrigir eventuais imóveis mal precificados, como explicado na seção 3.6.

Assim, para a validação do resultado do filtro de ocupação avaliou-se manualmente cerca de 100 imóveis que tiveram suas categorias modificadas pelo filtro, tanto aumentando quanto diminuindo-os de categoria. A seguir, destaca-se dois exemplos (um imóvel que teve sua categoria aumentada e um imóvel que teve sua categoria diminuída) que elucidam o funcionamento do filtro de taxa de ocupação:

Figura 12 – Exemplo de Imóvel TOP que teve sua categoria aumentada para MASTER. Fonte (fotos): Airbnb

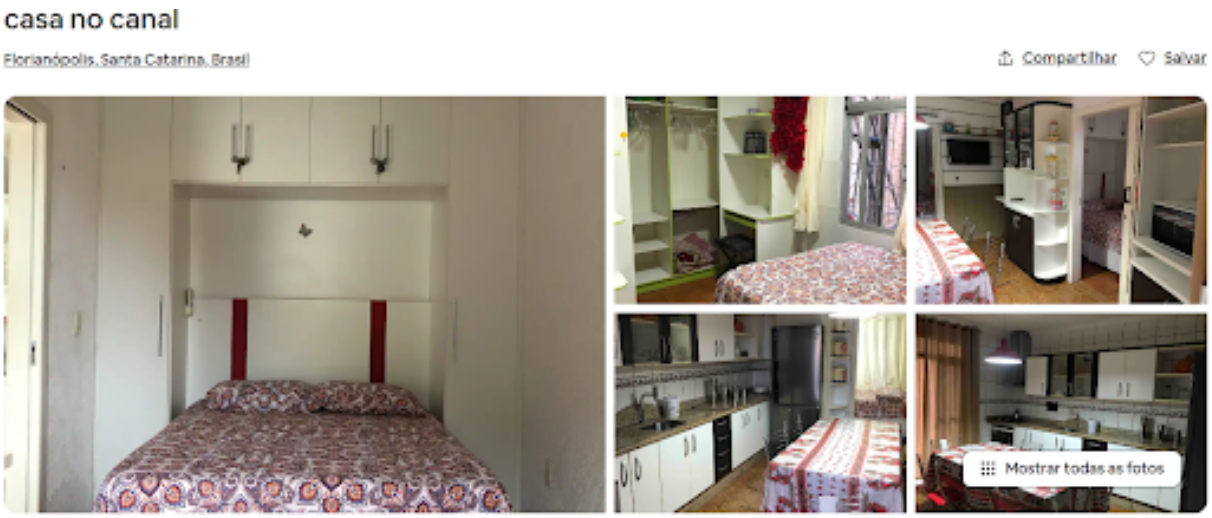


Figura 13 – Exemplo de Imóvel TOP que teve sua categoria diminuída para SUP. Fonte (fotos): Airbnb

Como explicado na seção 3.6, o imóvel que teve sua categoria rebaixada teve uma ocupação muito menor do que imóveis do mesmo cenário, indicando que o preço da diária estaria acima do esperado para este imóvel. Analogamente, o imóvel que teve sua categoria aumentada teve uma ocupação muito maior do que imóveis do mesmo cenário, indicando que o preço da diária estaria abaixo do esperado para este imóvel. Este fenômeno está comprovado na tabela 4, onde nota-se que o filtro foi aplicado nos imóveis seguindo o cálculo proposto nas equações 2 e 3:

Listing ID	Oc. Média (real)	Oc. Média (modelo)	Desvio Padrão	Resultado do Filtro
41264505	3%	38,7%	30,2%	-1 Categoria
11519407	95%	63,4%	30,2%	+1 Categoria

Tabela 4 – Resultado do Filtro de Taxa de Ocupação para os dois imóveis apresentados.

4.1.3 3° MVP - Exemplos de imóveis categorizados no Airbnb

Após a validação da regressão quantílica para alocação dos imóveis em faixas de preço e do filtro de taxa de ocupação, é necessário avaliar se estes resultados são coerentes. Para isso, o modelo foi recalculado utilizando os dados de um ano completo e, após sua finalização, avaliou-se manualmente cerca de 100 imóveis categorizados pelo algoritmo, para verificar se as categorias resultantes seriam coerentes com o que se espera de imóveis de suas respectivas categorias. A seguir, destaca-se cinco exemplos (um de cada categoria) que elucidam parcialmente este processo de validação para apartamentos de dois quartos em Jurerê:

Figura 14 – Exemplo de Imóvel da categoria Simples (SIM).
Fonte: Airbnb

Figura 15 – Exemplo de Imóvel da categoria Júnior (JR).
Fonte: Airbnb

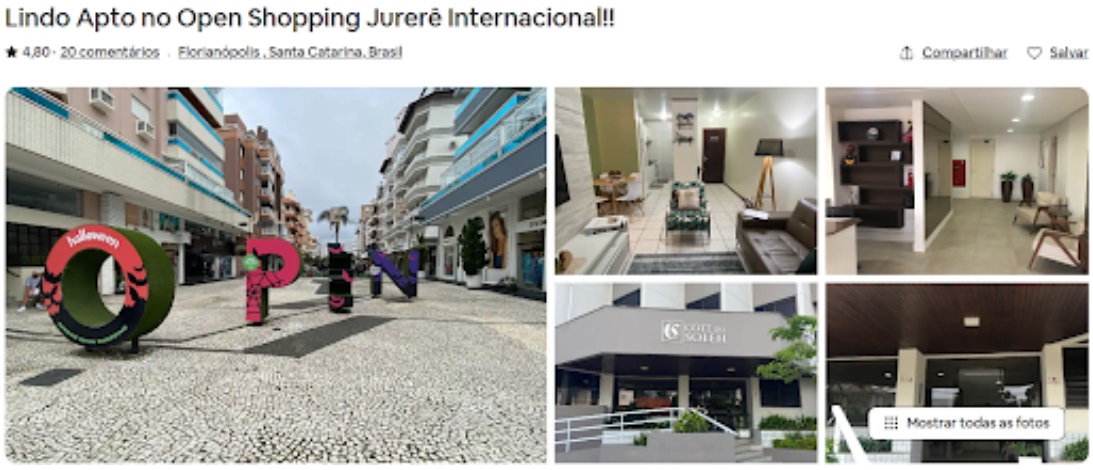


Figura 16 – Exemplo de Imóvel da categoria Superior (SUP).
Fonte: Airbnb

Figura 17 – Exemplo de Imóvel da categoria Top (TOP).
Fonte: Airbnb

Figura 18 – Exemplo de Imóvel da categoria Master (MASTER).
Fonte: Airbnb

Categoria do Imóvel	Preço Médio de Diária	Efeito do Filtro
Simples (SIM)	R$203,20	-
Júnior (JR)	R$340,38	-
Superior (SUP)	R$318,02	+1 Categoria
Top (TOP)	R$503,12	-
Master (MASTER)	R$881,60	-

Tabela 5 – Análise do preço médio de diária e efeito do filtro de taxa de ocupação *versus* categoria alocada para os imóveis selecionados.

A escolha deste cenário específico como destaque da análise é por justamente a empresa ter um conhecimento grande desta localidade e deste tipo de imóvel.

O imóvel alocado na categoria Simples (SIM) é inclusive da própria Seazone e já era considerado como um imóvel Simples internamente. Os imóveis das categorias Júnior e Superior (JR e SUP) apresentaram qualidade típica de imóveis destas categorias, em suas mobílias, tamanho e localidade e configurando uma categorização satisfatória para ambos os casos. Já o imóvel Top (TOP) apresentou uma qualidade maior de mobília e de localidade, além de apresentar maior tamanho e outras características que tendem a valorizar o imóvel, como um bom condomínio, piscina, varanda e vista-mar, configurando uma categorização satisfatória para este caso. Por fim, o imóvel Master (MASTER) apresenta características típicas de um imóvel Master, como ser um imóvel de cobertura, com jacuzzi e vista-mar, além da alta qualidade de forma geral em todos os aspectos do imóvel, configurando uma categorização também adequada.

Ao analisar os preços médios de diária temos que o modelo de regressão quantílica alocou os imóveis em suas respectivas categorias adequadamente, no entanto, nota-se que o imóvel Superior teve um preço de diária ainda menor do que o imóvel Júnior, mas isso se deu porque a princípio ambos seriam alocados na mesma categoria (Júnior), porém a ocupação deste imóvel foi muito alta se comparada a imóveis do mesmo cenário e categoria, o que fez com que o filtro atuasse aumentando sua categoria. Esta análise está evidenciada na tabela 5.

4.1.4 Modelo Final para todos os listings e Análise Geral

Dada a validação do modelo de categorização para os primeiros MVP's, calcula-se o modelo final utilizando todos os dados disponíveis no banco de dados. Assim, levando em consideração as variáveis de entrada definidas em 3.2, a base de dados e os procedimentos de limpeza e tratamentos dos dados definidos em 3.3, tem-se o resultado final, composto pela categorização de:

- 132.761 *listings* categorizados.
- Com 701 cidades avaliadas e computadas.
- Com 2396 bairros avaliados e computados.

O resultado final está parcialmente explícito na figura 19, evidenciando a saída do modelo separada por cenário e contendo uma categoria atribuída para cada *listing*.

	airbnb_listing_id	city	suburb	type	bedrooms	strata
1	31353538	Águas da Prata	Águas da Prata	apartament	2	JR
2	40263201	Águas da Prata	Águas da Prata	casa	2	MASTER
3	42528166	Águas da Prata	Águas da Prata	casa	2	SIM
4	43504641	Águas de Lindóia	Águas de Lindóia	casa	3	SIM
5	31855725	Águas de Lindóia	Águas de Lindóia	casa	4	SUP
6	42412219	Águas de Lindóia	Águas de Lindóia	hotel	1	SUP
7	17699872	Águas de Lindóia	Assunção	apartament	1	TOP
8	33807582	Águas de Lindóia	Centro	apartament	0	MASTER
9	36556862	Águas de Lindóia	Centro	apartament	1	SUP
10	50825927	Águas de Lindóia	Centro	apartament	1	SUP
11	23333004	Águas de Lindóia	Centro	apartament	2	SUP
12	40396417	Águas de Lindóia	Centro	apartament	2	SIM
13	42699983	Águas de Lindóia	Centro	apartament	3	MASTER
14	44155687	Águas de Lindóia	Centro	casa	2	SUP
15	31860804	Águas de Lindóia	Centro	casa	5	SUP
16	17187326	Águas de Lindóia	Centro	hotel	1	SUP
17	37551735	Águas de Lindóia	Jardim Lázari	casa	1	SIM
18	41691023	Águas de Lindóia	Jardim Lázari	casa	2	MASTER
19	31948260	Águas de Lindóia	Jardim Lázari	casa	3	SUP
20	39751545	Águas de Lindóia	Jardim Lázari	casa	4	TOP
21	52598254	Águas de Lindóia	Jardim Paraíso	apartament	1	SUP
22	48146320	Águas de Lindóia	Jardim Paraíso	casa	2	JR
23	37517090	Águas de Santa Bárt	Águas de Santa Bárbara	casa	2	JR
24	52436487	Águas de Santa Bárt	Águas de Santa Bárbara	casa	2	SUP
25	49591995	Águas de Santa Bárt	Águas de Santa Bárbara	casa	3	MASTER
26	46199631	Águas de São Pedro	Águas de São Pedro	apartament	1	JR
27	17680913	Águas de São Pedro	Águas de São Pedro	apartament	1	MASTER
28	31057983	Águas de São Pedro	Águas de São Pedro	apartament	1	JR
29	37092134	Águas de São Pedro	Águas de São Pedro	apartament	1	SIM

Figura 19 – Resultado final do Modelo de Categorização de imóveis.

Uma das principais dificuldades ao longo do processo de desenvolvimento e conclusão deste modelo se dá no caráter *Big Data* deste projeto. A otimização do código e das configurações do *Spark Cluster* são fundamentais para o processamento de uma quantidade grande de dados, como este trabalho demanda. De qualquer forma, atingiu-se o objetivo com um tempo de execução do algoritmo de aproximadamente três horas.

4.2 MODELOS DE PREDIÇÃO DE TAXA DE OCUPAÇÃO E FATURAMENTO

A seguir, são apresentados os resultados dos modelos de predição de taxa de ocupação e faturamento dos imóveis. Os resultados estão separados entre uma análise geral, uma análise comparativa com a base de dados disponibilizados pela empresa AirDNA e uma análise comparativa com os faturamentos reais conhecidos da própria Seazone.

4.2.1 Análise Geral e Dashboard de Faturamentos e Ocupações

Os modelos de predição da taxa de ocupação e faturamento dos imóveis também são calculados separados por cenário, seguindo as diretrizes de cenário definidas na seção 3.1.

Para o modelo de predição da taxa de ocupação, emprega-se o uso da regressão logística para obter uma porcentagem representativa a frequência de reservas de grupos de imóveis de um determinado cenário, como explicado na seção 3.5.

Já para o modelo de predição de faturamento, aproveita-se o modelo de regressão quantílica utilizado na categorização dos imóveis definindo quantis intermediários entre as faixas de preço de cada categoria, de forma a obter um valor que seja representativo a um preço médio de diária de cada categoria para cada cenário e juntamente com o resultado da taxa de ocupação provável calculado no modelo de predição de taxa de ocupação obtém-se uma predição de faturamento por cenário. Estes procedimento e cálculo estão explícitos em mais detalhes na seção 3.5 e na equação 1.

A execução do modelo completo para todos os cenários teve como resultado:

- 292.771 cenários calculados com predições de taxa de ocupação e faturamento.
- Predições por localização, para cada cidade e bairro do Brasil das quais temos dados.
- Predições por sazonalidade, para cada mês do ano.
- Predições por tipo de imóvel (apartamento, casa e hotel).
- Predições por número de quartos.

	city	suburb	month	type	bedrooms	strata	avg occ prob	revenue
1	Florianópolis	Aeroporto	1	apartamento	1	TOP	0.6314072499	3,466.4258017143
2	Florianópolis	Aeroporto	1	apartamento	1	MASTER	0.6746529738	5,892.787065795
3	Florianópolis	Aeroporto	1	apartamento	1	SUP	0.6304612929	2,619.2616100093
4	Florianópolis	Aeroporto	1	apartamento	1	SIM	0.6451918532	1,575.2607246388
5	Florianópolis	Aeroporto	1	apartamento	2	SIM	0.6098488185	2,425.6244941135
6	Florianópolis	Aeroporto	1	apartamento	2	SUP	0.6098488185	4,128.8732264764
7	Florianópolis	Aeroporto	1	casa	0	JR	0.8066284582	2,906.887306245
8	Florianópolis	Aeroporto	1	casa	1	MASTER	0.7850405224	7,114.0114745922
9	Florianópolis	Aeroporto	1	casa	1	JR	0.7832791856	4,335.4502922498
10	Florianópolis	Aeroporto	1	casa	1	TOP	0.7814258644	6,293.3197569454
11	Florianópolis	Aeroporto	1	casa	1	SIM	0.7814011439	2,372.1826340102
12	Florianópolis	Aeroporto	1	casa	2	MASTER	0.7660507502	11,409.4932605107
13	Florianópolis	Aeroporto	1	casa	2	TOP	0.7695507426	10,246.249901784
14	Florianópolis	Aeroporto	1	casa	2	SUP	0.7848431782	8,601.4327509055
15	Florianópolis	Aeroporto	1	casa	2	SIM	0.7660507502	3,814.932736043
16	Florianópolis	Aeroporto	1	casa	3	SIM	0.7482643723	6,060.9414160252
17	Florianópolis	Aeroporto	1	casa	3	MASTER	0.7967388981	19,680.0732359225
18	Florianópolis	Aeroporto	1	casa	3	TOP	0.7477493354	16,329.9271975629
19	Florianópolis	Aeroporto	1	casa	3	JR	0.7495975572	10,206.1319926189
20	Florianópolis	Aeroporto	1	casa	3	SUP	0.751246508	13,344.5313339754
21	Florianópolis	Aeroporto	1	casa	4	MASTER	0.7432488302	31,034.8857986501
22	Florianópolis	Aeroporto	1	casa	4	JR	0.7352147262	15,764.0540356009
23	Florianópolis	Aeroporto	1	casa	6	MASTER	0.6815474033	39,566.3223697799
24	Florianópolis	Aeroporto	1	casa	6	SIM	0.6979712665	25,598.0961991847
25	Florianópolis	Aeroporto	1	casa	6	TOP	0.6975238746	38,628.9240085806
26	Florianópolis	Aeroporto	2	apartamento	1	SIM	0.5030914561	1,041.3993141481
27	Florianópolis	Aeroporto	2	apartamento	1	TOP	0.4781235716	1,979.4315862656
28	Florianópolis	Aeroporto	2	apartamento	1	SUP	0.5041852125	1,368.8628519646
29	Florianópolis	Aeroporto	2	apartamento	2	SIM	0.4826018917	1,722.8887532651
30	Florianópolis	Aeroporto	2	apartamento	2	SUP	0.4826018917	2,128.2743422687
31	Florianópolis	Aeroporto	2	casa	1	JR	0.6786525079	3,804.332058673

Figura 20 – Resultado final dos Modelos de Predição de Taxa de Ocupação e Faturamento dos imóveis.

Além do registro dos resultados em uma tabela no banco de dados da empresa, há também a criação de um *dashboard* utilizando a ferramenta Power BI, descrita na seção 2.7. Este *dashboard* - presente na figura 21 - contém todas as predições de taxa de ocupação e faturamento e todos os imóveis categorizados separados por cada cenário. Ou seja, resume e condensa os resultados dos três modelos desenvolvidos neste trabalho em uma interface de fácil uso e acesso para o restante da empresa, permitindo que outros setores tenham acesso a essas informações e possam filtrá-las por cidade, bairro, tipo do imóvel, número de quartos e categoria do imóvel (cenário).

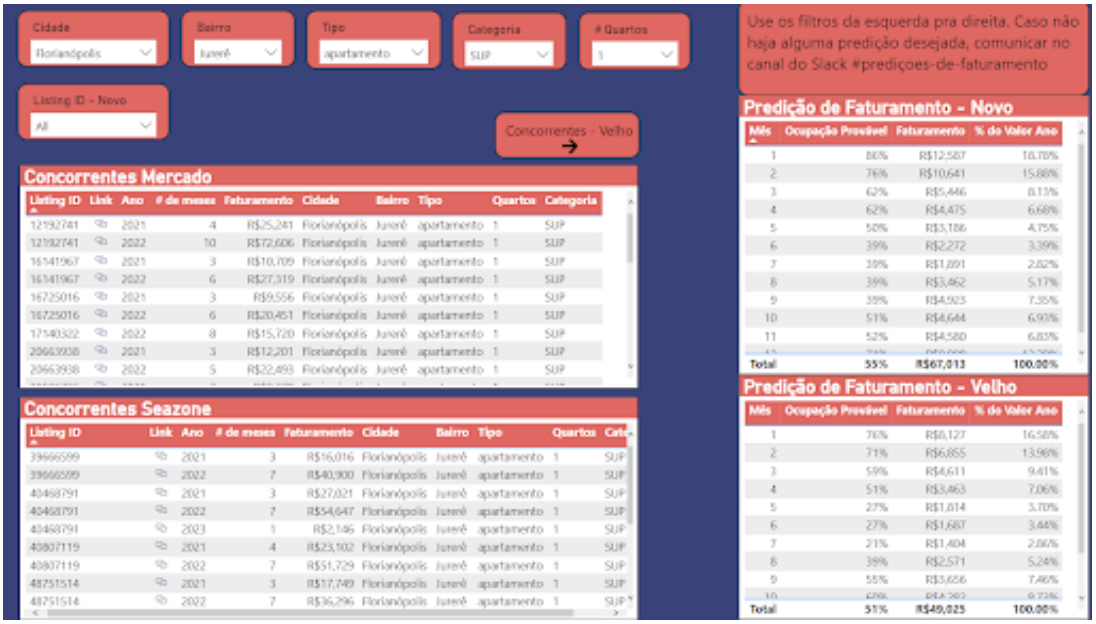


Figura 21 – Dashboard em Power BI com imóveis categorizados e predições de taxa de ocupação e faturamento por cenário.

4.2.2 Análise Comparativa com dados do AirDNA

Além do registro e disponibilização dos dados de saída do modelo, foram realizadas análises de validação destes resultados. Entre elas, comparou-se os resultados de faturamento do modelo com a base de dados comprada da empresa AirDNA. No entanto, vale ressaltar que o modelo produzido neste trabalho é calculado utilizando dados a partir de Setembro de 2021, enquanto que a base de dados utilizada na comparação conta com dados apenas do ano de 2021.

Como 2021 foi um ano de maior influência da pandemia no mercado de aluguel por temporada, houve também um maior impacto nos faturamentos dos imóveis de Airbnb nesta época.

É importante levar tudo isso em consideração ao observar alguns resultados como:

1. Disparidade de faturamento nos meses de alta temporada (como Janeiro e Fevereiro).

2. *Offset* de faturamento em alguns casos.

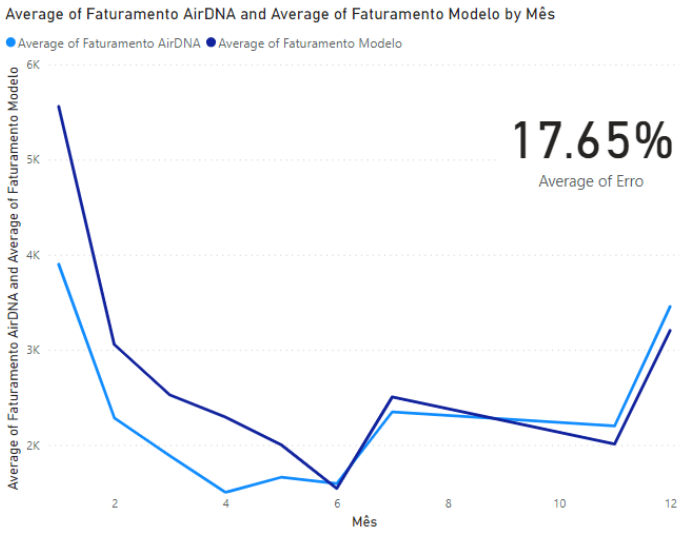


Figura 22 – Comparação da predição de faturamento calculada pelo modelo *versus* base de dados AirDNA para imóveis de Porto Seguro.

No gráfico de Porto Seguro, nota-se que tanto a sazonalidade quanto os valores de faturamento são capturados de forma satisfatória, contando com o módulo do erro absoluto de 17,65%. Mesmo assim, é notável a disparidade de faturamento nos meses de alta temporada, como comentado em 4.2.2.

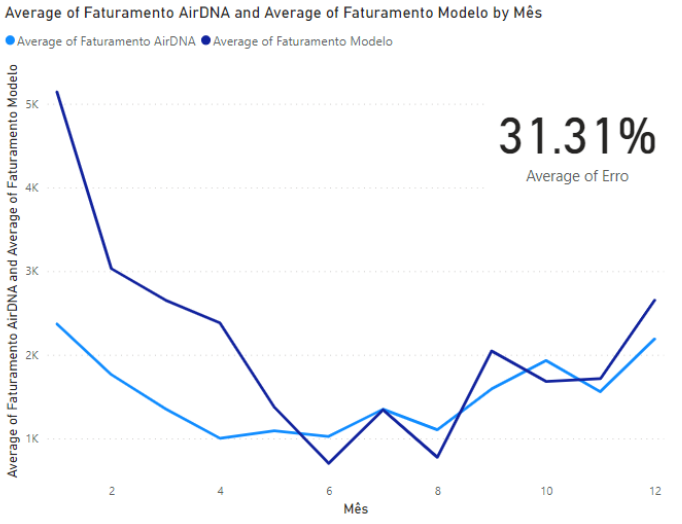


Figura 23 – Comparação da predição de faturamento calculada pelo modelo *versus* base de dados AirDNA para imóveis de Cabo Frio.

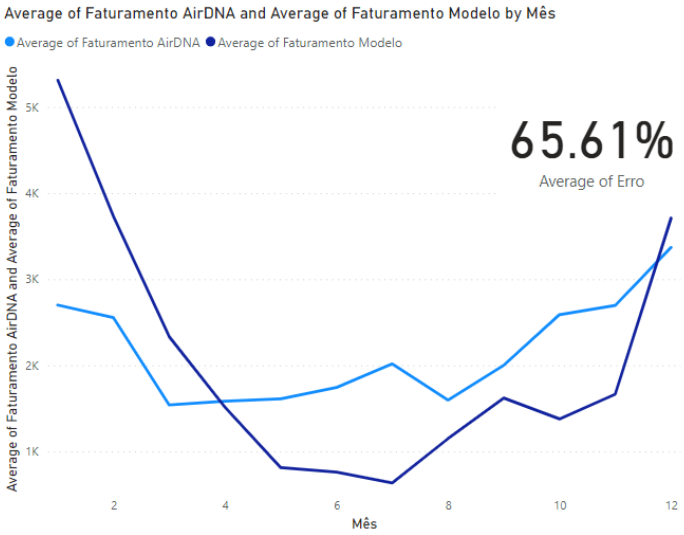


Figura 24 – Comparação da predição de faturamento calculada pelo modelo *versus* base de dados AirDNA para imóveis de Florianópolis.

Já nos gráficos de Cabo Frio e Florianópolis, nota-se que a sazonalidade é capturada satisfatoriamente, no entanto, há uma maior disparidade no valor dos faturamentos ao longo dos meses e principalmente nos meses de alta temporada, com módulo dos erros absolutos atingindo 31,31% e 65,61%, respectivamente.

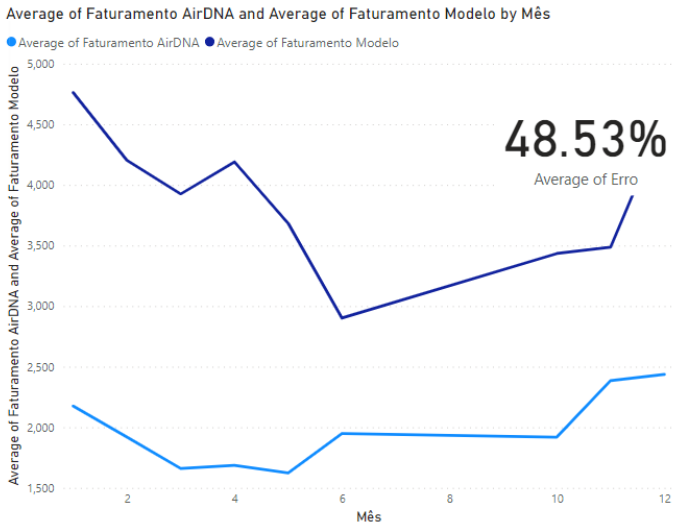


Figura 25 – Comparação da predição de faturamento calculada pelo modelo *versus* base de dados AirDNA para imóveis de Niterói.

Já no caso de Niterói, observa-se que a sazonalidade é capturada satisfatoriamente, no entanto, o fenômeno de *offset* de faturamento descrito em 4.2.2 é nítido, com módulo dos erros absolutos atingindo um valor médio de 48,53%.

Vale ressaltar que a representação da sazonalidade realmente é um ponto forte nos resultados deste modelo. Nota-se que em localidades litorâneas, onde há maior demanda por imóveis em meses quentes de alta temporada do ano, como Dezembro,

Janeiro e Fevereiro, o modelo é capaz de representar bem estas dinâmicas. Ao mesmo tempo, em localidades em que a demanda é maior nos meses mais frios do ano (como Junho e Julho), como em regiões de montanha, o modelo também é capaz de representar esta demanda, como nos gráficos 26 e 27 de faturamento e ocupação de apartamentos de um quarto da categoria SUP em Campos do Jordão (SP) e Gramado (RS), respectivamente.

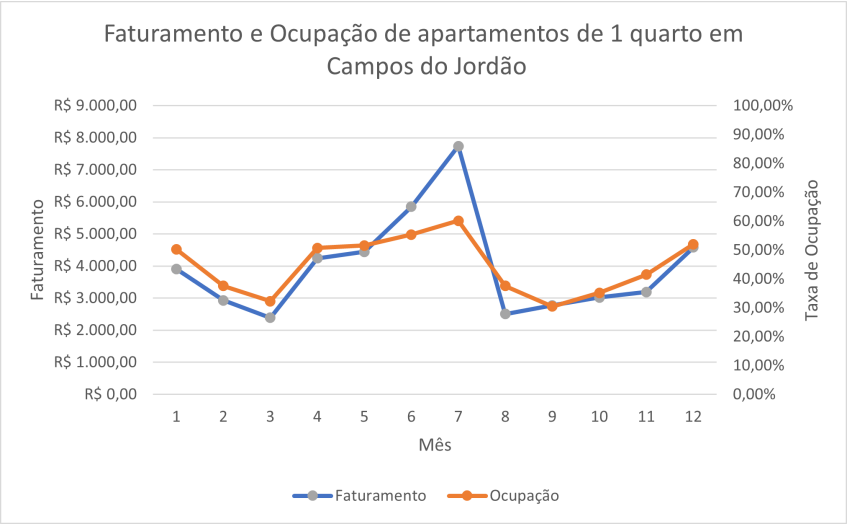


Figura 26 – Predições de faturamento e taxa de ocupação para apartamentos SUP de um quarto na cidade de Campos do Jordão.

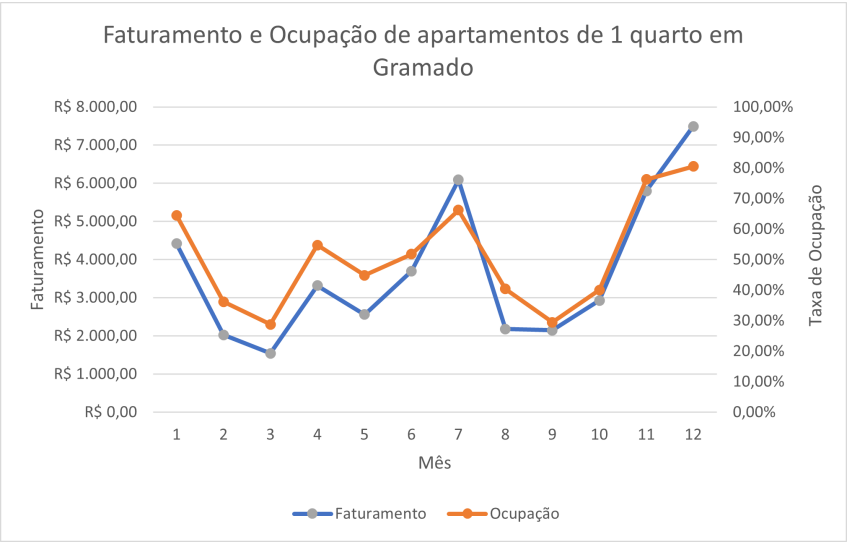


Figura 27 – Predições de faturamento e taxa de ocupação para apartamentos SUP de 1 quarto na cidade de Gramado.

4.2.3 Análise Comparativa com dados da Seazone

Por fim, realizou-se uma análise comparativa com o faturamento dos *listings* em administração pela própria Seazone. Esta análise contém as comparações dos

faturamentos *listing* a *listing*, por categoria e por região, também separadas mensal, trimestral, semestral e anualmente. Os resultados estão dispostos na tabela 6.

Listing	Mensal	45,33%
	Trimestral	41,74%
	Semestral	42,35%
	Anual	30,84%
Categoria	Mensal	48,95%
	Trimestral	43,78%
	Semestral	43,28%
	Anual	30,80%
Região	Mensal	48,05%
	Trimestral	42,45%
	Semestral	42,99%
	Anual	30,48%

Tabela 6 – Resultado da análise de erro do modelo *versus listings* da Seazone.

É importante ressaltar alguns pontos que podem influenciar nos resultados de erro.

A categorização dos imóveis da Seazone foi realizada manualmente, observando apenas as fotos dos imóveis, independente da localização do imóvel e da comparação com *listings* similares do cenário a qual esse *listing* se encontra. Já o modelo atribui uma categoria com base na faixa de qualidade e preço de imóveis disponíveis para um dado cenário. Isto faz com que em um cenário onde há muitos imóveis de diferentes qualidades, é possível observar uma maior distinção entre imóveis de diferentes categorias, porém em cenários em que há, por exemplo, poucos imóveis de baixa qualidade há pouca distinção entre imóveis de diferentes categorias. Nestes casos, é possível que um imóvel categorizado como Master, por exemplo, não aparente ser um imóvel que se encaixe nesta categoria, mas para este cenário específico seria realmente o imóvel de maior qualidade e ofertado pelo maior preço, comparativamente a outros imóveis do mesmo cenário. Isto quer dizer que ao realizar a comparação entre categorias, há uma distinção prática entre a definição de categoria para os imóveis da Seazone e para os imóveis do modelo.

Além disso, ao se comparar o faturamento *listing* a *listing* é importante ressaltar que existe um fator de aleatoriedade de aproximadamente 30% nas reservas e, portanto, no faturamento dos *listings*. Isto provém de um estudo efetuado internamente na própria Seazone em que os faturamentos e reservas de aproximadamente 50 imóveis praticamente idênticos situados em uma mesma localização foram avaliados. Ou seja, um erro por *listing* de aproximadamente 30% é considerado satisfatório. Desta forma, temos que os resultados anuais obtidos são considerados satisfatórios pela empresa, no entanto, os resultados mensais, trimestrais e semestrais poderiam ser melhorados.

5 CONCLUSÃO

Em conclusão, temos que o trabalho aqui realizado propôs o desenvolvimento de três modelos baseados em dados dos imóveis do Airbnb do Brasil: um modelo de categorização dos imóveis que recebe uma lista de *listings* e suas respectivas variáveis de entrada e retorna uma categoria, condizente com a qualidade e características do imóvel, para cada *listing*, um modelo de predição da taxa de ocupação destes imóveis e um modelo de predição de faturamento destes imóveis.

Para o modelo de predição da taxa de ocupação, utilizou-se o método da regressão logística como principal algoritmo estatístico para o cálculo da ocupação, baseando-se nos dados de eventos de reserva (binário) de cada imóvel para cada dia do ano. Já para os modelos de categorização e predição de faturamento utilizou-se a regressão quantílica como principal algoritmo estatístico, sendo que para o primeiro o modelo utiliza-se de quantis de preço (logarítmico) para definir faixas de preço de diária e, ao alocar os imóveis nestas faixas, define-se uma categoria entre cinco possíveis. Para o segundo, utiliza-se quantis médios de preço (intermediários aos quantis que definem as faixas de preço de cada categoria) de forma a obter um preço médio correspondente a grupos de imóveis separados por cenário e por categoria e, juntamente à predição da taxa de ocupação também obtida via modelo, calcula-se a predição de faturamento dos imóveis.

Outro desafio enfrentado se dá no caráter *Big Data* deste trabalho, já que os dados processados atingem a ordem de centenas de *gigabytes* de memória em tabelas e bilhão de registros de diárias (cada registro corresponde a uma linha na tabela). Para isso, utilizou-se o PySpark, uma tecnologia de computação distribuída em nuvem para se trabalhar com *clusters* de dezenas de máquinas processando dados e executando cálculos paralelamente em tempo real.

Com isso, foi possível categorizar 132.761 *listings* e prover a predição de taxa de ocupação e faturamento de 292.771 cenários, separados por localização, sazonalidade, tipo do imóvel e número de quartos, tudo isso com tempo de execução de três horas e utilizando um *cluster* com 20 máquinas trabalhando em paralelo. O custo de execução dos três modelos na AWS é de aproximadamente US$15,00 e poderá ser otimizado ainda mais no futuro.

Além disso, com erros comparativos absolutos na ordem de 30%, os resultados foram considerados majoritariamente satisfatórios e foram rapidamente aproveitados e utilizados por outros setores da empresa, como o setor Comercial e o setor de Investimentos, permitindo que o processo de expansão da empresa fosse melhor direcionado para regiões de maior faturamento e ocupação, fornecendo maior entendimento do mercado de aluguel por temporada para cada localidade do Brasil, permitindo saber qual é o melhor tipo de imóvel e que características trazem maior rentabilidade para

diferentes localidades, proporcionando maior capacidade de realizar melhores análises com grupos de imóveis separados por categoria e proporcionando uma maior capacidade de captação e retenção de clientes demonstrando domínio e conhecimento por parte da empresa sobre o mercado em diferentes localidades de atuação.

Não menos importante, automatizou-se o processo de categorização de imóveis, o qual antes era realizado manualmente, consumindo tempo e recursos humanos neste processo, principalmente ao considerar que foram mais de 100 mil *listings* categorizados de uma só vez. Além disso, a categorização fornecida pelo algoritmo mostrou-se coerente e robusta, fornecendo resultados personalizados para cada cenário.

Mesmo assim, discutiu-se alguns pontos que poderiam ser potenciais melhorias para estes modelos no futuro, entre eles:

- Utilizar as coordenadas dos imóveis e um raio de proximidade como parâmetro de entrada para o modelo, em contrapartida a utilizar a cidade e o bairro do imóvel. Desta forma seria possível capturar peculiaridades de cada cidade que não são captadas apenas utilizando o bairro como granularidade mínima de localização.

- Adicionar novas variáveis de entrada para os modelos de categorização e predição, como avaliações no Airbnb, número de comentários, a presença e quantidade de *ammenities* como ar condicionado, vista-mar, jacuzzi, entre outros fatores que podem influenciar na categoria, ocupação e faturamento do imóvel.

- Utilizar outros métodos de aprendizado de máquina, como *Random Forests* ou redes neurais, que podem trazer maior precisão e robustez na obtenção tanto dos resultados de categorização quanto de predição, se comparados a métodos como as regressões quantílica e logística.

REFERÊNCIAS

DASGUPTAA, S.; DEB, U. K. Binary Logistic Regression Models for short term prediction of premonsoon convective developments over Kolkata (India). *In*: DEPARTMENT OF STATISTICS, ST. XAVIER'S COLLEGE, 30 PARK STREET, KOLKATA 700016, INDIA ATMOSPHERIC SCIENCE RESEARCH GROUP, SCHOOL OF ENVIRONMENTAL STUDIES, JADAVPUR UNIVERSITY, KOLKATA 700032, INDIA. SCI-HUB – INTERNATIONAL JOURNAL OF CLIMATOLOGY. [*S.l.*: *s.n.*], 2007. Disponível em: https://sci-hub.se/10.1002/joc.1449.

GERACI, Marco. Modelling and estimation of nonlinear quantile regression with clustered data. *In*: ARNOLD SCHOOL OF PUBLIC HEALTH, DEPARTMENT OF EPIDEMIOLOGY e BIOSTATISTICS, UNIVERSITY OF SOUTH CAROLINA, COLUMBIA SC, USA. ELSEVIER – Computational Statistics and Data Analysis 136 (2019) 30–46. [*S.l.*: *s.n.*], 2019. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167947318302846.

GRAH, Simon. **6 recommendations for optimizing a Spark job**. [*S.l.*: *s.n.*], nov. 2021. Disponível em: https://towardsdatascience.com/6-recommendations-for-optimizing-a-spark-job-5899ec269b4b.

HANSEN, Karsten T. **Predicting Binary Events**. [*S.l.*], 2021. Disponível em: https://bar.rady.ucsd.edu/bin_class.html.

KARAGOZ, Sercan. **Creating Apache Spark Standalone Cluster with on Windows**. [*S.l.*: *s.n.*], jan. 2021. Disponível em: https://medium.com/analytics-vidhya/creating-apache-spark-standalone-cluster-with-on-windows-95e66e00a2d8.

KIRKOS, Efstathios. Airbnb listings' performance: determinants and predictive models. *In*: VARNA UNIVERSITY OF MANAGEMENT. KIRKOS (2022) – / European Journal of Tourism Research30, 3012. [*S.l.*: *s.n.*], 2022. Disponível em: :ejtr.vumk.eu/index.php/about/article/view/2142/518.

MUSSA, Adriano; YANG, Edward; TROVÃO, Ricardo; FAMÁ, Rubens. Hipótese de Mercados Eficientes e Finanças Comportamentais – As discussões Persistem. *In*: PONTIFÍCIA UNIVERSIDADE DE SÃO PAULO – PUC-SP. SEGET – Simpósio de Excelência em Gestão e Tecnologia. [*S.l.*: *s.n.*], 2004. Disponível em: https://www.aedb.br/seget/arquivos/artigos07/1241_TextoSeget.pdf.

SERVIÇO PÚBLICO DO ESPÍRITO SANTO, ESESP - Escola de. **Introdução ao BPM e Modelagem com BMPN 2.0**. [*S.l.*: *s.n.*], 2019. Disponível em: https://esesp.es.gov.br/Media/esesp/Apostilas/APOSTILA%20-%20CURSO%20ESESP%20BPM-%20BPMN.pdf.

VILLARREAL, Marco. **GitHub: Docker Spark Cluster**. [*S.l.*: *s.n.*], 2021. Disponível em: https://github.com/mvillarrealb/docker-spark-cluster.