UNIVERSIDADE FEDERAL DE MINAS GERAIS

MSI I - RELATÓRIO FINAL - MODELO PARA PREDIÇÃO DE ALOCAÇÃO DE MORADIAS (*HOME PLANNING*)

Modelo de predição para alocação de moradia, que visa desenvolver um sistema baseado em aprendizado de máquina capaz de prever a moradia mais adequada para um determinado indivíduo, com base em dados como renda, situação, motivo, entre outros.

Pesquisa Mista
Nome: Fernando Eduardo Pinto Moreira
Matrícula: 2019054536
Orientador: Adriano Alonso Veloso

Belo Horizonte
2023

1. Introdução

A alocação de moradia é uma decisão importante na vida de muitas pessoas, envolvendo diversos fatores como renda, situação, motivo, localização, entre outros. Neste contexto, a utilização de técnicas de aprendizado de máquina para criar um modelo de predição pode trazer benefícios significativos.

Observa-se também o impacto significativo dos benefícios habitacionais oferecidos pelo governo, que beneficiam centenas de milhares de moradias em todo o país. De acordo com dados do governo federal, entre 2019 e 2022, aproximadamente 1,6 milhão de moradias foram entregues por meio do Ministério do Desenvolvimento Regional (MDR).

O objetivo deste projeto é desenvolver um modelo capaz de prever qual a moradia mais adequada para uma pessoa com base em dados pré-fornecidos. Este modelo de predição tem o potencial de otimizar o processo de alocação de imóveis e melhorar a experiência do cliente, fazendo-o alcançar o seu objetivo de adquirir uma morada.

Com base nisso, este projeto visa a construção de um modelo de aprendizado de máquina, para prever moradias de programas sociais do governo federal e outras moradas, utilizando dados reais, como nome, motivo, presença de filhos, renda, entre outros. Estes dados são divididos em dados de entrada ou X, os quais são usados juntamente com um dado de saída y, que representa o resultado - no caso deste problema, y possui o valor de 0 se a pessoa não irá passar daquela determinada etapa ou 1, caso contrário. Este par (X, y) em conjunto formam o dado. Como temos 7 etapas ao todo, teremos 7 colunas que representam o y.

Com base nisso, o modelo será capaz de identificar a partir desses dados as melhores opções de moradia para cada cliente, auxiliar na solução de pendências e possíveis problemas, otimizar o potencial de crédito e encontrar um plano de aquisição do imóvel. No âmbito desse projeto, será realizado o tratamento adequado dos dados, visando facilitar sua utilização no modelo.

Além disso, serão extraídas informações e padrões relevantes a partir dos dados coletados, permitindo a classificação dos clientes em diferentes etapas de um funil composto por 7 etapas. Essas etapas representam os estágios nos quais os clientes se encontram com base em suas informações e situações específicas.

Dessa forma, o objetivo geral é construir um modelo de aprendizado de máquina, abordando desde conceitos fundamentais até sua aplicação prática, para facilitar a alocação de moradias aos clientes. Além disso, os objetivos específicos incluem o tratamento adequado dos dados, a extração de informações relevantes, a classificação dos clientes em etapas do funil e a alocação das moradias mais adequadas a cada perfil de cliente. Estas etapas serão explicadas mais detalhadamente ao longo deste relatório.

2. Referencial Teórico

2. 1. Aprendizado de Máquina

O aprendizado de máquina ou *machine learning* é um campo da inteligência artificial que se concentra no desenvolvimento de algoritmos e técnicas capazes de permitir que as máquinas aprendam a partir de dados, sem serem explicitamente programadas. É uma área ampla que abrange diferentes tipos de aprendizado, como o supervisionado, não supervisionado e por reforço. No contexto deste projeto, o aprendizado de máquina supervisionado foi utilizado com a técnica de classificação, pois temos dados rotulados para treinar o modelo e prever a moradia mais adequada com base nesses rótulos.

2.2. Pré-processamento de Dados

O pré-processamento de dados é uma etapa muito importante para garantir a qualidade e a adequação dos dados utilizados no treinamento do modelo. Essa etapa envolve a limpeza dos dados, tratamento de valores faltantes, codificação de variáveis categóricas, normalização de dados numéricos, entre outros. O pré-processamento dos dados é importante para remover ruídos e inconsistências que possam prejudicar o desempenho do modelo. Foi bastante utilizado neste projeto para um melhor desempenho dos algoritmos.

2.3. Algoritmos de Aprendizado de Máquina

Neste projeto, alguns algoritmos de aprendizado de máquina foram utilizados, como o Naive Bayes, Máquinas de Vetores de Suporte (SVM), Regressão Logística, k-Nearest Neighbors (KNN), Árvores de Decisão, Random Forest e Extreme Gradient Boosting (XGBoost). Abaixo segue uma breve explicação de cada um.

2.4. Naive Bayes

O Naive Bayes é um algoritmo de aprendizado de máquina supervisionado baseado no Teorema de Bayes. Ele assume que as características são independentes entre si, dada a classe. Este algoritmo calcula a probabilidade de pertencer a cada classe com base nas probabilidades condicionais das características e seleciona a classe com a maior probabilidade. É rápido, simples e eficiente para conjuntos de dados grandes.

2.5. Support Vector Machines (SVM)

O SVM é um algoritmo que busca encontrar o hiperplano que melhor separa as classes no espaço de características. Ele maximiza a margem entre as classes, tornando-se robusto a dados ruidosos.

2.6. Regressão Logística

Este algoritmo é utilizado para problemas de classificação binária, a regressão logística estima a probabilidade de um evento ocorrer. É um modelo probabilístico que ajusta os coeficientes das variáveis independentes para prever a classe de uma nova instância.

2.7. k-Nearest Neighbors (KNN)

O KNN (k-Nearest Neighbors) é um algoritmo de aprendizado de máquina utilizado em problemas de classificação e regressão. Ele classifica uma nova instância com base nas classes dos k vizinhos mais próximos no espaço de características. O valor de k é um parâmetro que determina a quantidade de vizinhos considerados. O KNN utiliza a distância entre as instâncias para encontrar os vizinhos mais próximos e atribuir a classe mais frequente como a classe da nova instância.

2.8. Árvores de Decisão

Este algoritmo cria uma árvore de decisão a partir dos dados de treinamento, onde cada nó representa uma decisão baseada em uma variável. É uma técnica intuitiva e facilmente interpretável, que permite a geração de regras de decisão claras.

2.9. Random Forest:

O *Random Forest* é um algoritmo que cria um conjunto de árvores de decisão independentes e combina suas previsões para obter uma classificação final mais precisa. É um método eficaz para reduzir o overfitting e lidar com dados desbalanceados.

2.10. Extreme Gradient Boosting (XGBoost)

O XGBoost (Extreme Gradient Boosting) é uma biblioteca de aprendizado de máquina que utiliza o algoritmo de boosting para construir modelos preditivos de alta performance. Ele combina várias árvores de decisão fracas para criar um modelo forte. O XGBoost otimiza uma função de perda através de um processo iterativo, ajustando os pesos dos exemplos de treinamento para melhorar gradualmente o desempenho do modelo.

2.11. Avaliação de Modelos

A avaliação do desempenho do modelo é essencial para garantir sua eficácia e generalização para novos dados. Existem várias métricas de avaliação de modelos que podem ser aplicadas dependendo do tipo de problema e dos objetivos do projeto. Algumas métricas comuns incluem:

2.12. Acurácia

Mede a proporção de instâncias corretamente classificadas em relação ao total de instâncias. É uma métrica geralmente utilizada em problemas de classificação balanceados, onde todas as classes têm aproximadamente o mesmo número de exemplos.

2.13. Recall

Também conhecido como taxa de verdadeiros positivos, mede a proporção de instâncias positivas corretamente classificadas em relação ao total de instâncias positivas. É uma métrica importante quando o objetivo é minimizar os falsos negativos.

2.14. F1-score

É uma medida de equilíbrio entre precisão e recall, calculada como a média harmônica entre essas duas métricas. É útil quando as classes estão desequilibradas e não queremos dar mais importância a uma métrica em detrimento da outra. Além das métricas de avaliação, também é comum utilizar técnicas como validação cruzada, que divide o conjunto de dados em conjuntos de treinamento e teste, permitindo uma avaliação mais robusta do desempenho do modelo.

2.15. Aplicações na Área Imobiliária

O uso de técnicas de aprendizado de máquina na área imobiliária se mostrou promissor, permitindo a otimização de processos, a melhoria da experiência do cliente e a tomada de decisões mais assertivas. A alocação de moradia é um problema complexo que envolve múltiplas variáveis e preferências individuais. A aplicação de modelos de predição pode auxiliar na seleção de moradias que atendam melhor às necessidades e preferências dos clientes, considerando fatores como renda, tamanho da família, localização, entre outros.

2.16. Ética e Privacidade dos Dados

No desenvolvimento do modelo, levou-se em consideração questões éticas e de privacidade dos dados, garantindo a confidencialidade e a segurança das informações pessoais dos indivíduos envolvidos nos dados utilizados.

2.17. Overfitting

O overfitting ocorre quando um modelo de aprendizado de máquina se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados. Isso acontece quando o modelo captura os padrões e o ruído específico dos dados de treinamento, em vez de aprender os padrões gerais. O overfitting leva a uma baixa capacidade de generalização e pode resultar em um desempenho inferior quando aplicado a novos dados.

3. Atividades Realizadas

O trabalho então, foi dividido em algumas partes. Primeiramente, foi definido o problema e como resolvê-lo. Chegou-se à conclusão que a melhor abordagem para o problema seria a utilização do aprendizado supervisionado com a classificação. Com isso, as principais atividades realizadas foram:

3.1. Coleta e preparação dos dados

Primeiramente, os dados foram obtidos e passaram por um processo de preparação. Foi então inserido no código e preparado para iniciar uma análise exploratória.

3.2. Análise exploratória dos dados

Nesta etapa, realizou-se uma análise para compreender as características dos dados e identificar possíveis padrões e relações entre as variáveis. Esta etapa foi fundamental para o entendimento do conjunto de dados e para a definição das estratégias de pré-processamento e seleção de características. Para isso, foram impressas várias colunas, a quantidade de repetição de valores em cada uma, quantos valores diferentes existiam, para que houvesse uma noção maior sobre cada coluna do conjunto.

3.3. Pré-processamento dos dados

Com base na análise exploratória, foram selecionadas as variáveis mais relevantes para o modelo. Foi realizado o tratamento de valores faltantes, realizando normalização e transformações necessárias para garantir a qualidade dos dados utilizados no treinamento do modelo. Aqui foi onde foi dedicada a maior parte do tempo, uma vez que o conjunto de dados continha uma quantidade de dados NaN (isto é, dados não-existentes), vários dados em formato de texto que tiveram que ser tratados para serem transformados em dados numéricos e assim, obtermos um maior desempenho nos algoritmos, além da exclusão de colunas que não agregavam valor pois possuíam todos os valores vazios.

3.4. Treinamento do modelo

Nesta etapa, foram escolhidos algoritmos de aprendizado de máquina adequados para a tarefa de predição da moradia mais adequada. Diversos modelos foram treinados e avaliados, sendo ajustados os hiperparâmetros e realizado validação cruzada para avaliar o desempenho dos modelos. Foram utilizados algoritmos como Naive Bayes, Árvores de Decisão, Regressão Logística, Support Vector Machines (SVM), Random Forest e XGBoost. Os dados foram divididos em um conjunto de treino e um conjunto de teste, para que pudéssemos testar os algoritmos para novas entradas. Percebeu-se que a acurácia variava muito pouco nos dados de treino e nos dados de teste, o que nos mostra que o modelo obteve sucesso e não apresentou o *overfitting,* problema clássico em aprendizado de máquina.

4. Resultados Obtidos

Os resultados obtidos demonstraram que o modelo de predição desenvolvido apresentou uma alta capacidade de acerto na classificação da moradia mais adequada para as pessoas analisadas. Observou-se que a maioria dos algoritmos obteve acurácia acima de 95%. A tabela 1 a seguir mostra a acurácia obtida em cada etapa sobre alguns algoritmos. Outros algoritmos ainda estão sendo treinados, seguindo o cronograma apresentado no primeiro relatório.

	Algoritmos		
Etapa	Naive Bayes	SVM	KNN
2	99,57%	99,11%	96,34%
3	93,27%	96,40%	93,03%
5	99,89%	95,84%	95,20%
5.2	99,87%	98,31%	97,41%
6	99,90%	99,89%	98,80%
7	99,91%	99,73%	99,74%
8	100,00%	100,00%	99,97%

Tabela 1: algoritmos e seus desempenhos em cada etapa.

Assim, este projeto visa obter alguns benefícios para os clientes, como:

- Melhora da experiência do cliente: ao prever qual é a moradia mais adequada para uma determinada pessoa, é possível oferecer opções que melhor atendam às suas necessidades e desejos, melhorando a experiência do cliente e aumentando a probabilidade de uma maior satisfação.
- Redução do tempo de procura: ao prever em qual etapa do funil a pessoa estará, é possível ajudá-la na solução de um determinado problema e, assim, facilitar e melhorar a sua experiência para conseguir a moradia.
- Aumento da precisão das previsões: ao utilizar técnicas de aprendizado de máquina, é possível obter previsões mais precisas e confiáveis do que as obtidas por meio de métodos tradicionais.

5. Conclusão

Com base no exposto, o projeto de criação de um modelo de predição para alocação de moradia ou *Home Planning* demonstrou ser uma ferramenta promissora para auxiliar no processo de alocação de imóveis. O modelo desenvolvido foi capaz de prever com precisão a moradia mais adequada para os indivíduos analisados, proporcionando benefícios tanto para os clientes quanto para a empresa.

Assim, com o projeto, pôde-se fixar vários conteúdos aprendidos durante todo o curso, principalmente das matérias de Aprendizado de Máquina, Estatística e matérias de programação em geral.

Com isso, a próxima etapa trata-se da finalização do treinamento em diferentes algoritmos, fazer uma análise mais profunda sobre quais usuários caem em quais etapas, para procurar entender e ajudar os clientes a passarem de cada etapa e, assim, conseguirem o imóvel.

6. Referenciais Bibliográficos

MITCHELL, T. M. Machine Learning. McGraw-Hill, New York, March 1997. ISBN 0070428077.

https://morada.com.vc

VELOSO, Adriano. Conteúdo da disciplina de Aprendizado de Máquina. Departamento de Ciência da Computação. Universidade Federal de Minas Gerais. 2023.

https://semiengineering.com/deep-learning-spreads/

7. Apêndices

Projeto completo:

https://drive.google.com/drive/folders/1TLFbflB74KJxhgwycWNdzPAwtFq-UEGP?usp=sharing

Tratamento dos dados:

https://colab.research.google.com/drive/16SFrbqQ_WsUqqDTRnO__qyiRkKn6AJtT?usp=sharing

Aplicação dos algoritmos:

https://colab.research.google.com/drive/1l4RydQ8vc9Uj6scyW9KD6vZC5jQEUV-h