Trabalho 2 - Introdução ao Aprendizado de Máquina

Luís Rafael Sena

Dezembro 2024

Contents

1 Introdução

Este relatório apresenta o desenvolvimento do Trabalho 2 da disciplina EEL891 - Introdução ao Aprendizado de Máquina, ofertada no período 2024-2. O objetivo deste trabalho foi criar um sistema preditivo que estimasse o preço de imóveis com base em suas características, utilizando técnicas de regressão multivariável.

O conjunto de dados utilizado continha informações detalhadas de imóveis, incluindo características estruturais (como área útil e número de quartos), localização (bairro) e elementos diferenciais (como piscina e vista para o mar). O modelo preditivo foi treinado em um conjunto de dados de 20.000 imóveis e avaliado em um conjunto oculto, com as métricas calculadas automaticamente pelo Kaggle.

Ferramentas como pandas, numpy e scikit-learn foram utilizadas para manipulação e modelagem dos dados. Este relatório detalha as etapas de pré-processamento, seleção de modelos e análise de resultados, discutindo as métricas de avaliação e o desempenho final.

2 Pré-Processamento dos Dados

O pré-processamento é fundamental para preparar os dados e garantir a precisão dos modelos. Foram adotadas as seguintes estratégias:

2.1 Tratamento de Dados Não Numéricos

As variáveis categóricas, como tipo de imóvel e bairro, foram convertidas para representações numéricas utilizando *one-hot encoding*. Essa técnica foi escolhida por sua capacidade de representar categorias sem introduzir hierar-

quias artificiais. Além disso, colunas como `diferenciais` foram inicialmente excluídas devido a sua baixa correlação com o preço do imóvel.

2.2 Tratamento de Dados Faltantes

Valores ausentes foram tratados com diferentes estratégias:

- **Dados categóricos:** Foram preenchidos com a mediana das categorias disponíveis, garantindo uma representação central.
- **Dados numéricos:** Foram preenchidos com zero, simplificando a imputação e reduzindo possíveis impactos negativos.

Essas abordagens se mostraram eficazes na padronização do conjunto de dados, evitando problemas de inconsistência.

2.3 Normalização

A normalização foi realizada com o *RobustScaler*, que é robusto a outliers e ideal para conjuntos de dados com grande variabilidade. Essa técnica garantiu que variáveis em diferentes escalas tivessem pesos equilibrados no treinamento.

2.4 Seleção de Características

Uma matriz de correlação foi utilizada para identificar variáveis redundantes e pouco relevantes. Como resultado, algumas colunas, como `diferenciais`, foram excluídas para reduzir a dimensionalidade e melhorar o desempenho computacional.

3 Modelos Testados

Diversos algoritmos de regressão foram explorados para identificar a melhor abordagem preditiva. As principais estratégias adotadas estão descritas abaixo:

3.1 Random Forest Regressor

O *Random Forest Regressor* foi selecionado devido à sua capacidade de lidar com variáveis heterogêneas e capturar interações complexas. Diferentes configurações foram testadas, com variações na profundidade máxima (até 4750) e no número de estimadores.

3.2 Regressão Linear

A Regressão Linear foi utilizada como modelo base para avaliar a simplicidade e interpretabilidade das previsões. Embora menos flexível, demonstrou ser uma referência útil para comparação.

3.3 K-Nearest Neighbors Regressor (KNN)

O KNN Regressor foi testado com diferentes valores de k, pesos baseados em distância e métricas de proximidade (como euclidiana e Manhattan). Essa abordagem é sensível à normalização, o que reforçou a importância do pré-processamento adequado.

4 Métricas de Avaliação

4.1 Divisão dos Dados

O conjunto de dados foi dividido em:

- **Treino:** 81% dos registros para treinamento inicial.
- **Teste:** 9% dos registros para validação intermediária.
- **Validação:** 10% dos registros para avaliação final.

4.2 Erro Relativo Médio Quadrático (RMSPE)

A métrica principal utilizada foi o *Root Mean Square Percentage Error* (RMSPE), que mede o erro percentual médio entre os valores previstos e os reais. É particularmente útil em problemas onde os valores absolutos variam amplamente.

5 Resultados e Discussão

5.1 Modelo Selecionado

O *Random Forest Regressor* com profundidade máxima de 4750 apresentou o melhor desempenho no conjunto de validação, destacando-se por sua capacidade de capturar interações complexas entre variáveis.

5.2 Impacto do Pré-Processamento

A normalização com *RobustScaler* e a imputação baseada em mediana e zero foram cruciais para garantir a qualidade do conjunto de dados, especialmente em relação a outliers e valores ausentes.

5.3 Desempenho no Kaggle

O modelo final obteve um erro RMSPE inferior a 0.3 no conjunto de teste do Kaggle, demonstrando sua eficácia em prever preços de imóveis com alta precisão.

6 Conclusão

Este trabalho mostrou a importância de técnicas robustas de pré-processamento e seleção de modelos para problemas de regressão multivariável. A utilização de bibliotecas como scikit-learn e pandas foi essencial para implementar um pipeline eficiente e escalável.