Modelos de regressão aplicados a tráfego de ônibus

Eduardo Martins[1], **Bruna Assunção**[2]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Niterói – RJ – Brasil

eduardomg@id.uff.br, brunaassuncao@id.uff.br

Abstract. *In this study, we analyze bus traffic patterns using artificial intelligence techniques. From a detailed dataset, we selected crucial variables like day of the week, time and direction of the line. We apply several machine processing algorithms to predict trip durations. The results, visualized in graphs, stand out as each model's celebrated accomplishments. This work aims to guide the optimization of public transport systems through AI. Supplementary materials are available on GitHub.*

Resumo. *Neste estudo, analisamos padrões de tráfego de ônibus utilizando técnicas de inteligência artificial. A partir de um conjunto de dados detalhado, selecionamos variáveis cruciais como dia da semana, horário e sentido da linha. Aplicamos diversos algoritmos de aprendizado de máquina para prever a duração das viagens. Os resultados, visualizados em gráficos, destacam as eficácias variadas de cada modelo. Este trabalho visa orientar a otimização de sistemas de transporte público através da IA. Materiais complementares estão disponíveis no GitHub.*

1. Introdução

Em metrópoles e cidades em rápido crescimento, o sistema de transporte público desempenha um papel crucial, influenciando tanto a qualidade de vida dos residentes quanto a dinâmica econômica local. A eficiência e confiabilidade desse sistema determinam, em grande medida, a facilidade com que as pessoas transitam, trabalham e interagem em tais ambientes urbanos. A cidade de Florianópolis, como muitas outras ao redor do mundo, depende de sua rede de ônibus para facilitar a mobilidade urbana. No entanto, com o crescimento urbano vem a complexidade, tornando cada vez mais desafiador prever e otimizar os tempos de viagem dos ônibus. Felizmente, vivemos em uma era onde a tecnologia tem o potencial de enfrentar esses desafios. A disponibilidade de grandes volumes de dados de transporte e os avanços na inteligência artificial (IA) oferecem oportunidades sem precedentes para refinar a forma como entendemos e abordamos as questões de trânsito [Tomé 2021].

Neste trabalho, empregamos diversas técnicas de aprendizado de máquina no conjunto de dados de tráfego de ônibus de Florianópolis. A cidade, embora distante de nossa localização atual, apresenta um ambiente intrigante para tais estudos devido à sua própria unicidade em termos de desafios de transporte. Nosso principal objetivo é criar um modelo de previsão detalhado que possa ser aplicado não apenas em Florianópolis, mas também adaptado e refinado para outras cidades. Vamos explorar uma variedade de variáveis – desde dias específicos da semana e horários do dia até detalhes mais granulares, como o sentido da linha – para antecipar com precisão a duração das viagens de

ônibus. Em última análise, pretendemos avaliar a eficácia de cada técnica de IA, identificar o modelo mais preciso e discutir como tal modelo pode ser usado para melhorar os sistemas de transporte em ambientes urbanos diversificados.

2. Metodologia

Neste estudo, adotamos uma abordagem sistemática para investigar e comparar a eficácia de diversos algoritmos de aprendizado de máquina na previsão de métricas de tráfego de ônibus. A metodologia seguiu as seguintes etapas:

1. Pré-processamento dos Dados: Inicialmente, trabalhamos com um conjunto de dados em formato CSV que continha múltiplas colunas, incluindo informações detalhadas sobre as linhas de ônibus, horários de operação, sentido da viagem, entre outras. Para tornar os dados mais gerenciáveis e relevantes para nossas análises:
    - Remoção de Outliers: Utilizando técnicas estatísticas robustas, identificamos e removemos outliers que poderiam distorcer os resultados dos modelos.
    - Transformação de Dados: A partir do dataset original, extraímos e transformamos colunas para obter um novo dataframe contendo as seguintes informações: DiaDaSemana, HorarioDia, DataHoraIni, Sentido, Linha, TempoViagemMinutos e TotalGiros.
2. Escolha dos Modelos de Aprendizado de Máquina: Escolhemos alguns modelos que em nossa pesquisa pareceram mais eficientes:
    - Modelos Lineares: LinearRegression, Ridge, Lasso e ElasticNet.
    - Modelos Baseados em Árvores: RandomForestRegressor, DecisionTreeRegressor e GradientBoostingRegressor.
    - Outros Modelos: SVR (Support Vector Regressor), KNeighborsRegressor e MLPRegressor (*Multi-layer Perceptron Regressor*).

    Para cada modelo, utilizamos os hiperparâmetros padrão, com exceção do MLPRegressor, onde especificamos um *max_iter* de 100000 para garantir a convergência do modelo.
3. Treinamento e Validação dos Modelos: Dividimos o conjunto de dados em conjuntos de treinamento e teste. Cada modelo foi treinado usando o conjunto de treinamento e, em seguida, validado no conjunto de teste. Para avaliar o desempenho de cada modelo, utilizamos métricas como o Erro Quadrático Médio (MSE), o Erro Médio Absoluto (MAE) e o coeficiente de determinação (R^2).
4. Comparação dos Modelos: Com base nas métricas obtidas para cada modelo no conjunto de teste, realizamos uma comparação direta para determinar quais modelos performaram melhor e por que.

Com esta metodologia, pretendemos proporcionar uma análise abrangente e objetiva da capacidade de diferentes algoritmos em prever métricas cruciais associadas ao tráfego de ônibus.

3. Processamento e transformação de dados

No processo de desenvolvimento, decidimos utilizar Python e suas bibliotecas que já nos eram conhecidas, como *pandas*, *sklearn* e *matplotlib* para a manipulação e preparação dos

dados para análise. Para armazenamento de código, utilizamos o github, que se encontra neste repositório.

O primeiro passo foi focado na extração de outliers. Estes, em um conjunto de dados de tráfego de ônibus, podem representar eventos raros que, se não tratados, podem distorcer os resultados dos modelos de regressão.

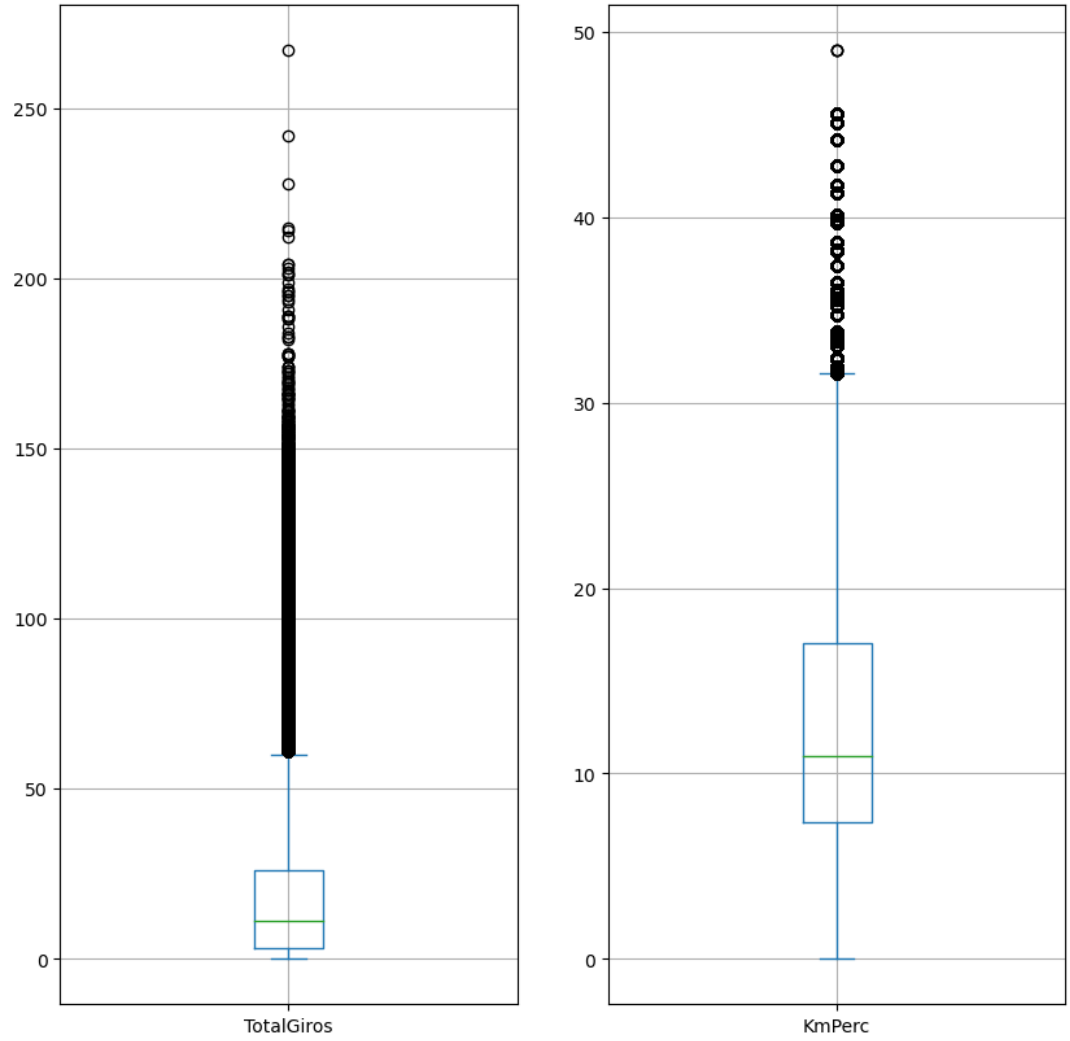

Figura 1. BoxPlot de Outliers

Uma análise realizada revelou a coluna "FimOperação" com uma considerável porcentagem (92%) de seus valores ausentes. Com esta informação, optamos por sua remoção para simplificar o conjunto de dados e assegurar que as variáveis mais significativas fossem enfatizadas.

A transformação dos dados foi uma etapa crucial. A partir das colunas originais, como "DataIni", "HoraIni", "DataFim", "HoraFim", "Sentido", "Linha", "NoVeículo", "DuraçãoViagem", "TotalGiros", "KmPerc", realizamos ajustes para criar um dataframe mais alinhado à análise de regressão. Geramos novas colunas, incluindo "DiaDaSemana", "HorarioDia", "DataHoraIni", "TempoViagemMinutos" e "TotalGiros", visando capturar aspectos essenciais do tráfego de ônibus que seriam pertinentes à previsão.

Com os dados devidamente preparados e estruturados, procedemos para a escolha dos algoritmos que utilizaríamos para a análise.

Os algoritmos foram escolhidos com base em nosso conhecimento prévio e popularidade, buscamos ter uma certa variedade de algoritmos entre regressores lineares, árvore de escolhas e alguns outros como MLPRegressor ou SVR, que foram adicionados por serem diferentes e com isso pudessem trazer um resultado interessante.

4. Resultados Obtidos

Nesta seção, apresentamos os resultados alcançados pelos diferentes algoritmos de aprendizado de máquina, avaliando a capacidade de cada modelo em prever métricas cruciais associadas ao tráfego de ônibus. Os modelos foram avaliados com base em três métricas-chave: RMSE (Erro Quadrático Médio Raiz), R^2 (Coeficiente de Determinação) e MAE (Erro Médio Absoluto). Para facilitar a compreensão dos resultados, os dados foram visualizados em três gráficos distintos, um para cada métrica.

1. RMSE (Erro Quadrático Médio Raiz): O RandomForestRegressor mostrou o desempenho mais promissor com um RMSE de 5.42, o que indica a menor quantidade de erro entre os modelos testados. No outro extremo, o MLPRegressor registrou o maior RMSE, chegando a 23.24, sugerindo uma previsão menos precisa.

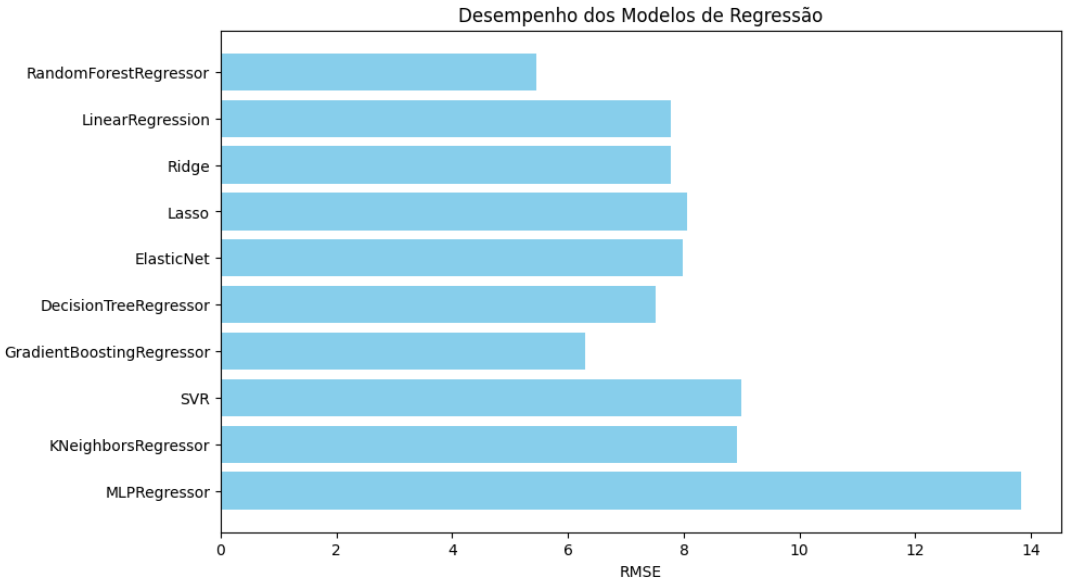

Figura 2. RMSE

2. R^2 (Coeficiente de Determinação): R^2 fornece uma medida da variação total dos resultados explicada pelo modelo. Um valor de R^2 próximo de 1 indica que o modelo pode explicar uma grande proporção da variância no resultado. Novamente, o RandomForestRegressor se destacou com um coeficiente R^2 de 0.60, indicando que o modelo foi capaz de explicar 60% da variabilidade nos dados de tráfego. Por outro lado, o SVR e KNeighborsRegressor apresentaram valores negativos, o que pode indicar que esses modelos são menos adequados para essa tarefa específica. O MLPRegressor, em particular, obteve um R^2 de -6.29, sugerindo um ajuste muito pobre ao conjunto de dados.

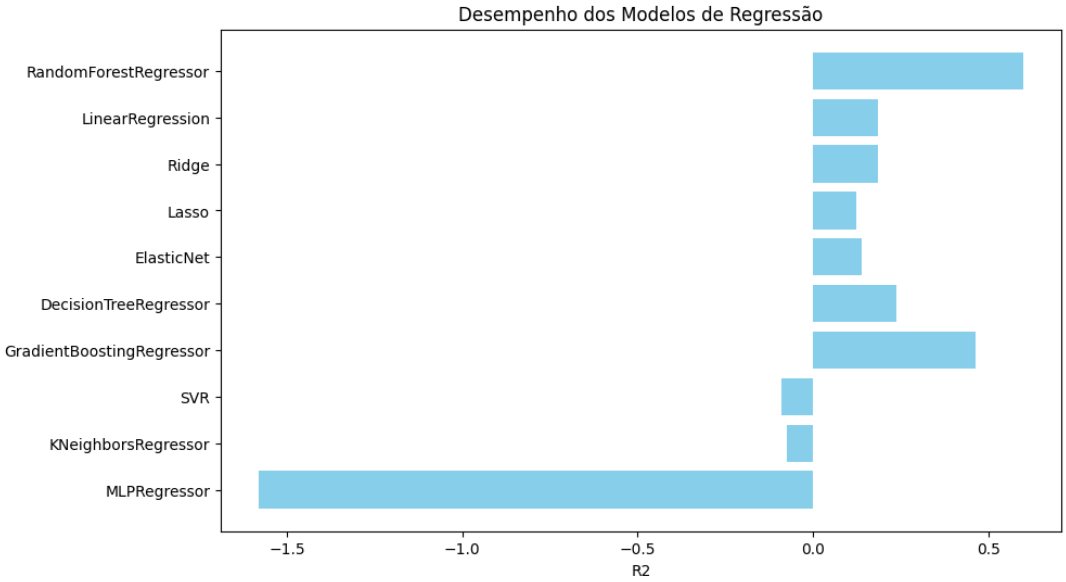

Figura 3. R^2

3. MAE (Erro Médio Absoluto): O MAE representa a diferença média entre as previsões e os valores reais. Um valor menor de MAE é desejável, pois indica previsões mais precisas. O RandomForestRegressor novamente liderou a lista com um MAE de 3.92, enquanto o MLPRegressor, com um MAE de 19.98, apresentou o maior erro em termos absolutos.

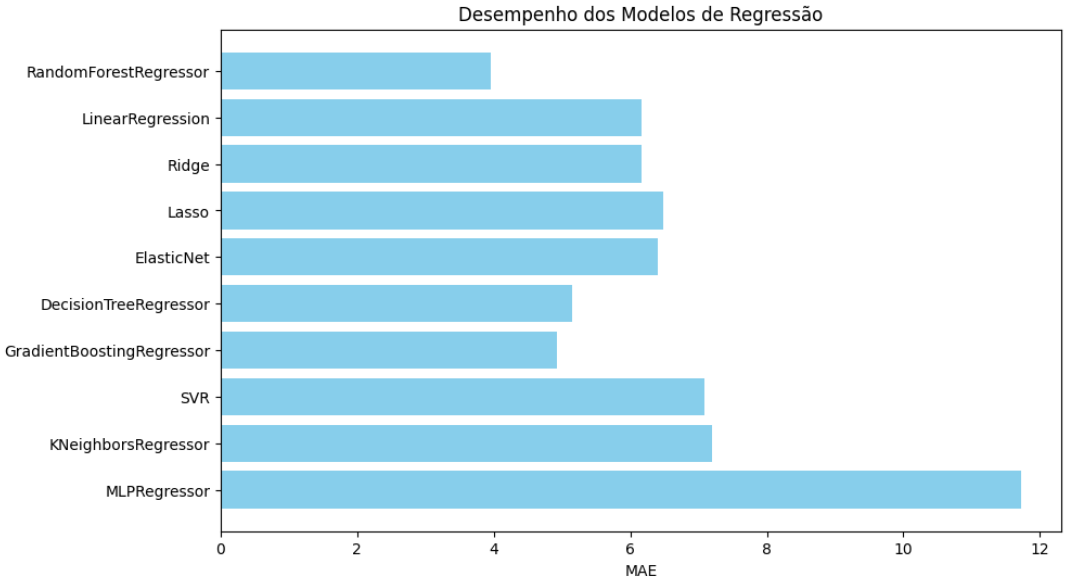

Figura 4. MSE

Conclusão Preliminar dos Resultados:

A partir dos dados apresentados, é evidente que o RandomForestRegressor apresentou um desempenho consistentemente superior em todas as três métricas, tornando-se o algoritmo mais promissor para a previsão de métricas de tráfego de ônibus neste estudo.

Por outro lado, o MLPRegressor mostrou um desempenho significativamente inferior, o que sugere a necessidade de mais ajustes ou reconsideração de seu uso para esta tarefa específica.

Os gráficos complementares a esta análise permitem uma visualização mais intuitiva das performances dos modelos, facilitando comparações diretas e uma melhor interpretação dos resultados obtidos.

5. Conclusões e Trabalhos Futuros

Ao avaliar uma série de algoritmos de aprendizado de máquina na tarefa de prever métricas de tráfego de ônibus, nossa pesquisa revelou insights significativos que podem ser benéficos para estudos futuros e aplicações práticas na área de transporte público.

1. Superioridade do RandomForestRegressor: O RandomForestRegressor destacou-se consistentemente em todas as métricas, evidenciando seu potencial como uma ferramenta robusta para tais previsões. Esse desempenho pode ser atribuído à capacidade do algoritmo de lidar com conjuntos de dados complexos e a sua natureza ensemble, que combina várias árvores de decisão para produzir previsões mais estáveis e precisas.
2. Cuidados com o MLPRegressor: O desempenho notavelmente inferior do MLPRegressor sugere que redes neurais, na configuração utilizada, podem não ser a melhor escolha para esta tarefa específica ou exigem um ajuste mais meticuloso de hiperparâmetros e arquitetura.
3. Importância da Pré-Processamento: A remoção de outliers e a transformação de dados foram etapas críticas que permitiram a extração de características significativas do dataset original. Isso reforça a ideia de que uma preparação cuidadosa dos dados é tão vital quanto a escolha do modelo.
4. Espaço para Otimização: Já que utilizamos a maior parte dos modelos com hiperparâmetros padrão, há um bom espaço para tentar buscar melhores parâmetros.

Os seguintes temas podem ser abordados no futuro:

- Otimização de Hiperparâmetros: Um estudo mais aprofundado focado na otimização de hiperparâmetros para cada modelo pode revelar melhorias de desempenho significativas.
- Inclusão de Mais Características: Considerando a complexidade do tráfego de ônibus, a introdução de outras características, como condições climáticas, feriados ou eventos locais, pode aprimorar a precisão das previsões.
- Modelos de Aprendizado Profundo: Modelos como LSTM ou GRU, que são especializados em sequências temporais, podem ser explorados, dada a natureza temporal dos dados de tráfego.

Referências

Tomé, P. T. F. (2021). Modelos de previsão de demanda: Uma aplicação no transporte rodoviário interestadual de passageiros por Ônibus na região sul do brasil. *Universidade Federal de Santa Catarina*.