Estimação de duração da viagem e número total de passageiros com uso de modelos de regressão e classificação.

Lucas Dirk Gomes Ferreira[1], Matheus Veloso da Silva[1], Miguel Domingos Brito[1]

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

{lucasdirk@id.uff.br, matheusveloso@id.uff.br, migueldb@id.uff.br}

Abstract: *This article is a study of the application of Machine Learning methods for the estimation of important variables in public transport via bus, namely: trip duration and total number of passengers. Simpler methods such as linear regression were used, to ensemble methods such as Random Forest and Gradient Boosting based on histograms. The analysis of the best models took into account the performance measures adopted and the computational cost, in order to propose an efficient solution with the lowest possible cost. Regarding performance measures, the most commonly used and easiest to understand measures were adopted in order to facilitate the understanding of our proposal.*

Resumo: *Este artigo é um estudo de aplicação de métodos de Machine Learning para a estimação de variáveis importante no transporte público via ônibus, a saber: duração da viagem e número total de passageiros. Foi utilizado de métodos mais simples como regressão linear, até métodos ensemble como Random Forest e o Gradient Boosting baseado em histogramas. A análise dos melhores modelos levou em consideração as medidas de desempenho adotadas e o custo computacional, com a finalidade de propor uma solução eficiente e de menor custo possível. Sobre as medidas de desempenho, foram adotadas as mais comumente utilizadas e de mais fácil compreensão a fim de facilitar o entendimento de nossa proposta.*

1. Introdução

Neste artigo abordaremos a construção de um modelo classificação usando a técnica Random Forest e modelos de regressão utilizando a regressão linear para aplicação em um problema de transporte público. O dataset utilizado neste artigo se encontra neste endereço do github.: https://github.com/TielleAlexandre/datasetFlorianopolis/blob/main/dataGPS_Floripa.zip. O endereço do github dos arquivos referentes ao presente artigo é: https://github.com/MigDB27/ML_Floripa.

O transporte público é parte essencial para o funcionamento de uma sociedade. A deslocação usando ônibus é a mais comum em qualquer cidade desenvolvida. No Brasil, este modo de transporte se torna ainda mais importante devido à ausência de metrôs e trens em muitas cidades, mesmo naquelas com população elevada, sendo presentes geralmente em capitais.

A gestão de empresas de ônibus tem sido um desafio de complexa solução. Um dos desafios está na estimativa da duração da viagem, essencial para determinar quantos ônibus irão rodar por dia, como organizar os horários dos funcionários, bem como os gastos associados. Os fatores que interferem na duração da viagem são variados como:

itinerários possuírem trajetos longos e com muitas pausas, pela variação de picos de utilização do transporte no decorrer do dia e também pela presença de problemas de excesso de tráfego nas estradas, especialmente devido a dependência alta deste tipo de transporte.

Para estimar tal informação crucial é necessário um modelo de aprendizado de máquina que leve em conta os fatores relacionados com a duração da viagem. Neste conjunto de dados, se encontram dados referentes as viagens das linhas de ônibus da cidade de Florianópolis, estando presente dados referentes: à data de partida, à data de saída, à hora de partida, à hora de saída, à linha que o ônibus pertencia, ao sentido da viagem, ao número do veículo, à duração da viagem, ao número giros da roleta na viagem e ao número de quilômetros percorridos.

No algoritmo de regressão, foi escolhido como atributo classe o atributo “Duração da Viagem” e o “Total de Giros”. No algoritmo de classificação também foi escolhido “Duração da Viagem” como atributo classe, porém este foi separado em faixas pela regra de Sturges, sendo assim um classificador com duas classes.

2. Referencial Teórico e Revisão da Literatura

Este artigo aborda técnicas de Machine Learning para desenvolver modelos de regressão e classificação. Para isso é necessário fazer o pré-processamento dos dados para então treinar os modelos.

No algoritmo de regressão foi utilizado a técnica de regressão linear. Esta é a técnica de regressão mais utilizada e consiste na predição de valores contínuos desconhecidos, a variável dependente, com base em outros valores de dados relacionados, as variáveis independentes. Essa relação se dá matematicamente a partir de uma equação linear.

Outra técnica muito importante usada foi a Regra de Sturges para separar valores numéricos em faixas. Esta regra é um método empírico proposto por Herbert Sturges, utilizado para determinar o número de classes que devem existir num histograma de frequência, com a finalidade de realizar a classificação dos dados de uma amostra ou de uma população. Para isso foi considerado um diagrama de frequência ideal de k intervalos, onde o i-ésimo intervalo contém o número de amostras dada pela equação abaixo:

$$C_{(k-1,i)} = C_i^{k-1} = \binom{k-1}{i}$$

Equação 1: Cálculo do número de amostras no intervalo pela Regra de Sturges

O número de amostra desta regra consiste no número de maneiras pelas quais podem ser usadas para obter um subconjunto do conjunto de dados relacionado. Isto é possível pelo coeficiente binomial dado pela equação abaixo:

$$N = \sum_{i=0}^{k-1} \binom{k-1}{i} = 1 + 1^{\,k-1} = 2^{k-1}$$

Equação 2: Número de amostras

Para simplificar a equação, Sturges utilizou a propriedade logarítmica e obteve a equação para calcular o número ideal de intervalos:

$$k = 1 + log_2(N)$$

Equação 3: Número ideal de intervalos

Para o problema de classificação a Regra de Sturges foi essencial para determinar o número de classes. O modelo de classificação escolhido para treinar um classificador foi o de Random Forest. Este algoritmo é baseado em outro algoritmo de aprendizado de máquina, a saber, o de árvores de decisão. O Random Forest gera aleatoriamente várias árvores de decisão e combina o resultado obtido em todas elas para fornecer um resultado final.

As árvores de decisão são algoritmos cuja estrutura é similar à de uma árvore, onde os ramos são os caminhos percorridos em direção as folhas que contém o resultado da predição.

Por ser um algoritmo do tipo ensemble, isto é, que combina resultado de algoritmos mais simples, este algoritmo requer um custo computacional maior, mas entrega melhores resultados, e seu custo é menor que algoritmos de redes neurais, sendo uma alternativa ao uso destes.

Uma das vantagens das florestas randômicas é que não costumam sofrer overfitting e não são muito afetados por outliers, além de possuírem uma boa performance em bases desbalanceadas. Considerando que outliers e desbalanceamento são comuns em muitas bases de dados, este modelo foi adotado.

Para avaliação do modelo foi adotado as métricas mais comumente utilizadas que são: acurácia, precisão, recall e f1-score.

A acurácia corresponde ao total de acertos pelo total de amostras. Essa métrica descreve o quanto o modelo acerto de uma maneira geral.

$$Acurácia = \frac{Verdadeiros\ Positivos\ (TP) + Verdadeiros\ Negativos\ (VN)}{Total}$$

Equação 4: Cálculo da acurácia

A precisão corresponde ao total de valores positivos classificados corretamente pelo total de valores classificados como positivos. Essa métrica descreve o quanto o modelo é capaz de evitar falsos positivos.

$$Precisão = \frac{Verdadeiros\ Positivos\ (TP)}{Verdadeiros\ Positivos\ (TP) + FalsosPositivos\ (FP)}$$

Equação 5: Cálculo da precisão

O recall ou revogação corresponde ao total de valores positivos classificados corretamente pelo total de valores que são positivos, incluindo os que foram classificados erroneamente como negativos. Essa métrica descreve o quanto o modelo é capaz de evitar falsos negativos.

$$Recall = \frac{Verdadeiros\ Positivos\ (TP)}{Verdadeiros\ Positivos\ (TP) + Falsos\ Negativos\ (FN)}$$

Equação 6: Cálculo do recall

E a métrica F1, que é uma medida balanceada do recall e da precisão. Descreve a qualidade geral da classificação.

$$F1 = \frac{2 * precisão * recall}{precisão + recall}$$

Equação 7: Cálculo do F1-score

Por fim, foi decidido experimentar outro modelo, devido aos resultados insatisfatórios, este modelo foi o modelo de regressão Gradient Boosting baseado em histogramas. Esse modelo também é da família ensemble, como todos modelos boosting. Os modelos Boosting são uma classe de algoritmos de aprendizado conjunto que adicionam modelos de árvore a um conjunto de forma sequencial. Entretanto, ao usar muitas árvores numa base com um número significativo de dados, isso gera uma lentidão do treinamento e custo computacional excessivo, como visto no classificador de Random Forest. Com a finalidade de otimizar o processo de treinamento um outro modelo Gradient Boosting baseado em histogramas foi selecionado.

Este algoritmo diminui o tempo utilizado para treinar a árvore de decisão ao reduzir o número de valores para recursos de entrada contínua convertendo os valores contínuos em faixas de um histograma, o reduz o número de valores exclusivos drasticamente, reduzindo o custo computacional. Essa simplificação tem pouco impacto no modelo, quando o impacto não é nulo, mas o ganho de economia nos recursos utilizados é muito alto. Importante ressaltar que a Regra de Sturges não foi utilizada visto que o próprio algoritmo já separa valores contínuos em histogramas.

Foi então decidido avaliar este modelo de regressão com outras métricas além do coeficiente de correlação devido ao seu bom resultado. Essas métricas foram a variância explicada, o erro médio absoluto, o erro quadrado médio e o erro mediano absoluto.

O coeficiente de correlação indica o quanto a reta se aproxima e consegue explicar a distribuição dos dados, a equação para efetuar o cálculo segue abaixo:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Equação 8: Cálculo do coeficiente de correlação R2

Sendo y o valor predito, $\hat{y}$ o valor esperado e $\bar{y}$ a média de y.

A variância explicada é o quanto de variância o modelo explica, números próximos de 1 indicam um bom modelo, e próximos de 0 um modelo incapaz de explicar a variância dos dados. A equação da variância explicada é descrita abaixo:

$$explained\_variance(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}}$$

Equação 9: Cálculo da variância explicada

Sendo Var a variância.

O erro médio absoluto é calculado a partir da média dos erros absolutos, ou seja, utilizamos o módulo de cada erro. Dessa forma valores negativos na diferença dos erros não afetam o cálculo. O cálculo é descrito pela equação abaixo:

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|$$

Equação 10: Erro médio absoluto

Sendo $n_{samples}$ o número de amostras.

O erro quadrado médio é calculado elevando ao quadrado a diferença dos erros o que leva a um maior peso aos maiores erros. Valores outliers afetam bastante o resultado desta métrica. A equação para cálculo segue abaixo:

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2.$$

Equação 11: Erro quadrado médio

E por último o erro mediano absoluto, que é calculado de forma similar ao erro médio absoluto, porém usa a mediana ao invés da média. Essa métrica é menos sensível a outliers. A equação segue abaixo:

$$\mathrm{MedAE}(y, \hat{y}) = \mathrm{median}(|\, y_1 - \hat{y}_1 \,|, \ldots, |\, y_n - \hat{y}_n \,|).$$

Equação 12: Erro mediano absoluto

3. Metodologia da Análise Experimental

A primeira etapa foi a de carregamento dos dados. A base inicialmente estava em formato de bando de dados Acess (.mdb), e foi convertida em uma planilha Excel (.xlsx). Para isso, cada tabela foi exportada para uma planilha Excel. O arquivo Excel foi armazenado no Google Drive para ser carregado no Jupyter Notebook no Google Colab, para gerar o dataframe. Neste estudo foram selecionados apenas os valores referentes a abril de 2019, devido ao fato de a base ser extensa o suficiente para análise do modelo.

Após criado o dataframe, a etapa de pré-processamento e análise exploratória tem início. No dataset estavam presentes valores onde continham viagens sem duração, essas linhas foram excluídas. As colunas relativas ao final da viagem foram excluídas bem como as relativas a data de início das viagens. Isso devido ao fato que já se possuía a duração da viagem.

Posteriormente, ao fazer a regressão linear foram excluídos outliers. Além da remoção destas linhas, foi gerada um atributo que continha o dia da semana que a viagem foi feita. Esses dias da semana foram convertidos em valores discretos. Outra tabela gerada foi a conversão do horário inicial em minutos.

A análise exploratória também foi feita com base no atributo classe duração da viagem, onde foram plotados histogramas para analisar a distribuição da variável e gráficos de dispersão dos outros atributos em relação aos atributos classe de duração da viagem. A distribuição dos dados relativos aos quilômetros percorridos, número do veículo, duração da viagem e total de giros da roleta foram feitas com histogramas.

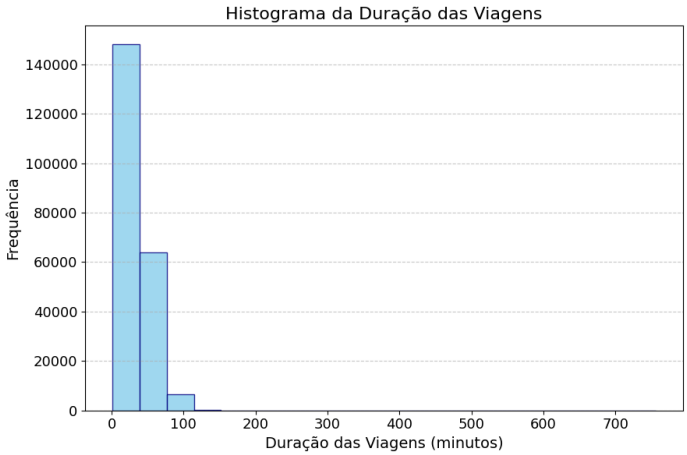

Gráfico 1: Histograma de frequência de duração das viagens

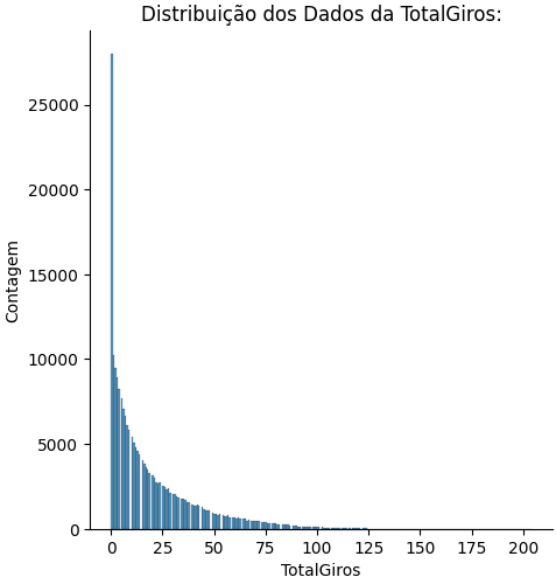

Gráfico 2: Histograma de frequência de duração das viagens

Para complementar a análise exploratória, uma matriz de correlação foi feita e com isso foi visto que as variáveis possuíam pouca correlação entre si, logo não deveriam ser retiradas. A duração da viagem foi convertida em faixas, essas faixas foram determinadas pela Regra de Sturges já previamente explicada na seção anterior deste artigo. A duração da viagem como valor contínuo foi utilizada na regressão linear e a convertida em faixas foi utilizada no modelo de classificação.

A próxima etapa foi o modelo de regressão linear. Após gerar o modelo foi feita análise dos resíduos com base em gráficos de dispersão e histogramas. Os outliers dos atributos referentes ao número total de giros da roleta e duração da viagem foram removidos e com isso pode ser executado o modelo. Dada a complexidade do problema, o resultado foi satisfatório, sendo este descrito na próxima seção.

Para o modelo de classificação foi utilizado o modelo Random Forest. A amostra foi separada em 70% de base de treino e 30% de bases de testes. A duração da viagem foi separada em duas faixas e o modelo obteve excelentes resultados com altíssima acurácia e valores de falsos positivos e negativos muito baixos.

Motivados pela boa performance do classificador, o modelo de regressão Gradient Boosting baseados em histogramas foi adotado, os motivos já foram anteriormente

discutidos. Os valores de duração da viagem e o total de giros da roleta foram preditos. O modelo apresentou um elevado coeficiente de correlação em ambos os casos. Ao predizer a duração da viagem o modelo se comportou muito bem, porém aconteceram erros significativos ao estimar o número total de giros. O modelo de Random Forest foi utilizado para estimar o número total de giros numa regressão e diminuiu os erros significativamente.

4. Resultados

Através da tabela 1, observamos que o modelo de regressão linear apresentou um coeficiente de determinação de 0.689, o que significa que cerca de 68.9% da variabilidade da variável 'Duração Viagem' é explicada pelas variáveis independentes incluídas no modelo. Isso indica um ajuste moderadamente bom do modelo aos dados, considerando-se a complexidade do fenômeno em estudo.

A estatística F apresentou um valor extremamente alto de aproximadamente 79690, indicando que o modelo como um todo tem uma relação linear significativa com a variável dependente "DuraçãoViagem". O valor-p associado à estatística F é igual a 0, o que indica uma significância estatística altamente relevante para o modelo.

Os coeficientes das variáveis independentes fornecem informações sobre suas contribuições individuais para a previsão do tempo de viagem. Notamos que as variáveis "Horaini", "Linha", "Sentido", "NoVeículo", "TotalGiros" e "KmPerc" possuem coeficientes significativos com valores menores que 0.05, o que significa que todas essas variáveis são relevantes para explicar as variações na variável resposta.

Com isso, o modelo estimado foi dado por:

$$y = 2.677 - 7.05e-05 * X_1 - 0.0010 * X_2 + 0.1559 * X_3 + 2{,}691e-6 * X_4 + 0.0125 * X_5 + 0.0711 * X_6$$

Equação 13: Equação da reta obtida na regressão linear

Onde, x1 = HoraIni, x2 = Linha, x3 = Sentido, x4 = NoVeículo, x5 = TotalGiros, x6 = KmPerc.

OLS Regression Results

Dep. Variable:	DuraçãoViagem	R-squared:	0.689
Model:	OLS	Adj. R-squared:	0.689
Method:	Least Squares	F-statistic:	7.969e+04
Date:	Tue, 18 Jul 2023	Prob (F-statistic):	0.00
Time:	04:20:18	Log-Likelihood:	-1.3775e+05
No. Observations:	215570	AIC:	2.755e+05
Df Residuals:	215563	BIC:	2.756e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.6777	0.004	651.413	0.000	2.670	2.686
HoraIni	7.05e-05	3.15e-06	22.349	0.000	6.43e-05	7.67e-05
Linha	-0.0010	1.87e-05	-52.470	0.000	-0.001	-0.001
Sentido	0.1559	0.002	76.847	0.000	0.152	0.160
NoVeículo	2.691e-06	4.87e-08	55.236	0.000	2.6e-06	2.79e-06
TotalGiros	0.0125	4.91e-05	253.756	0.000	0.012	0.013
KmPerc	0.0711	0.000	554.809	0.000	0.071	0.071

Omnibus:	18272.590	Durbin-Watson:	1.019
Prob(Omnibus):	0.000	Jarque-Bera (JB):	118531.473
Skew:	0.071	Prob(JB):	0.00
Kurtosis:	6.630	Cond. No.	1.19e+05

Tabela 1: Resultados da Regressão Linear

Para utilizar o modelo de classificação Random Forest, consideramos apenas três faixas derivadas da regra de Sturges: (1.0, 40.737], (40.737, 80.474], e (80.474, 120.211]. Essa seleção foi feita devido à pouca representatividade das outras faixas na amostra, sendo assim, elas foram removidas do conjunto de dados. Isso nos permitiu focar nas faixas que possuem maior relevância para o modelo e simplificar o processo de classificação, garantindo resultados mais significativos.

O modelo de classificação Random Forest obteve resultados muito expressivos, a-acurácia obtida foi de 93%.

As demais métricas que comprovam a eficácia do modelo e o pouco número de falsos negativos e positivos estão abaixo.

Classe	Precisão	Recall	F1-score
0	0.95	0.97	0.96
1	0.87	0.85	0.86
2	0.74	0.52	0.61

Tabela 2: Resultados do modelo de classificação usando Random Forest

Altos números de precisão, recall e por consequência, F1-score indicam baixa presença de valores falsos negativos e falsos positivos.

Para obter uma melhor estimativa da duração da viagem foi utilizado o Histogram-based Gradient Boosting. O modelo obteve bons resultados com uma alta correlação e valores de erros muito baixos, os resultados seguem na tabela abaixo:

Coeficiente de correlação	0,876
Variância explicada	0,876
Erro médio absoluto	0,199
Erro quadrado médio	0,084
Erro mediano absoluto	0,146

Tabela 3: Resultado do modelo de regressão usando Histogram-based Gradient Boosting para predição da duração da viagem

Pode-se dizer que os valores previstos apresentam poucos desvios da média e da mediana, e que não erros muito discrepantes de acordo com o erro quadrado médio.

O mesmo foi feito para predizer o número total de giros da catraca. Embora a correlação tenha sido alta, o erro foi significativamente maior, conforme o resulto na tabela abaixo:

Coeficiente de correlação	0,820
Variância explicada	0,820
Erro médio absoluto	5,984
Erro quadrado médio	82,626

Erro mediano absoluto	3,735

Tabela 4: Resultado do modelo de regressão usando Histogram-based Gradient Boosting para predição do número total de giros

Importante notar a diferença do erro quadrado médio, ou seja, há valores contendo um erro alto, embora a maioria não se afaste tanto da média. O erro mediano foi ainda menor, provando que a maioria não tende a se afastar tanto da mediana.

Para diminuir os erros foi treinado um modelo de regressão Random Forest, porque o modelo baseado em histogramas admite simplificações que podem impactar a predição. Os valores de correlação e variância explicada foram maiores e os erros menores, especialmente o erro quadrado médio.

Coeficiente de correlação	0,852
Variância explicada	0,852
Erro médio absoluto	5,170
Erro quadrado médio	67,868
Erro mediano absoluto	3,030

Tabela 5: Resultado do modelo de regressão Random Forest para predição do número total de giros

A redução foi significativa, devido ao fato do modelo Random Forest ser um modelo mais complexo. Outra questão a se considerar é a própria variável total de giros. O número de passageiros, intrinsicamente, é mais difícil de prever que a duração da viagem, isso porque fatores de decisões do indivíduo tem mais influência sobre esta variável. Logo, em se tratando de uma variável mais afetada pela escolha de apenas poucos indivíduos, o que não ocorre na duração da viagem, esta variável naturalmente apresentará maior erro, o que não impede a busca de um modelo que minimize mais os erros. Entretanto, em todos os modelos de regressão que seguem o método ensemble, a correlação foi alta, indicando que há correlação entre o atributo classe e os atributos previsores.

5. Conclusão e Trabalhos Futuros

Os resultados obtidos com os modelos foram satisfatórios e comprovam a eficácia do uso modelos preditivos de Machine Learning no problema em questão. A predição foi feita com pouco sucesso no modelo de regressão linear mais com muito sucesso nos modelos ensemble. O modelo de classificação teve um desempenho razoável.

A começar pela regressão linear, o modelo não obteve um resultado tão alto quando os dos modelos ensemble, o que pode indicar a relação entre os atributos previsores e o atributo classe não é linear. A distribuição dos resíduos não está dentro da normalidade, logo isso também afeta o modelo.

O modelo de classificação teve ótimo desempenho a descrever duas das três classes. Provavelmente, valores muito heterogêneos foram agrupados na classe com menos acertos. Outro ponto é que quanto maior o número de classes, maior a tendência a ter dificuldade na separação. O que se pode concluir é que a abordagem de classificação não é cabível neste problema dado ao fato de o modelo ter dificuldade em predizer poucas classes e devido a um desempenho bem melhor da abordagem de modelos de regressão.

Quanto aos modelos de regressão, estes foram tiveram um ótimo desempenho. Para a predição da duração da viagem, um modelo ensemble mais simples e rápido já conseguiu dar uma ótima solução, evitando custos computacionais e sendo possível aplicá-lo a bases maiores sem problemas.

A predição do número total de giros teve menos sucesso, e o modelo de Random Forest teve de ser utilizado. Ainda que o modelo apresente erros, esses não inviabilizam o uso deste modelo como um bom estimador, visto que a variável tende a ser mais difícil de ser prevista pois está sujeita a interferência de fatores humanos de forma mais sensível que a da duração da viagem. Entretanto, a busca por um melhor modelo se faz necessária, sendo, portanto, uma oportunidade de trabalho futuro. Uma observação é que não foi feito o tuning dos parâmetros, logo este pode ser um ponto de partida para melhora do modelo, ou mesmo, a adoção de modelos alternativos. A alta correlação encontrada justifica uma busca pela melhoria.

O artigo usou apenas os registros de um mês, e o estudo da aplicação deste modelo para os demais meses é outra oportunidade de trabalho futuro. Ou seja, analisar a validade desse modelo para descrever os registros dos outros meses, e em caso de problemas de performance, como melhorá-lo para que possa servir de um modelo preditor para qualquer época do ano

Este artigo pode ainda servir de base para aplicação de modelos similares a estimação do tempo de viagem em linhas de ônibus de outras cidades com diferentes realidades. Os excelentes resultados servem de ótimo indício para a viabilidade deste tipo de aplicação.

Outra aplicação possível seria usar uma metodologia similar para prever a duração de viagens para outros meios de transportes diferentes de ônibus e mesmo para viagens para que motoristas de carro, permitindo assim um estudo mais aprofundado do transporte público.

6. Referências Bibliográficas

https://maestrovirtuale.com/regra-de-sturges-explicacao-aplicacoes-e-exemplos/, último acesso em 17/07/2023

https://www.monolitonimbus.com.br/transformacao-box-cox/, último acesso em 17/07/2023

https://aws.amazon.com/pt/what-is/linear-regression/#:~:text=A%20regress%C3%A3o%20linear%20%C3%A9%20uma,independente%20como%20uma%20equa%C3%A7%C3%A3o%20linear, último acesso em 17/07/2023

https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems, último acesso em 17/07/2023

https://machinelearningmastery.com/histogram-based-gradient-boosting-ensembles/, último acesso em 17/07/2023

https://scikit-learn.org/stable/modules/classes.html#module-sklearn.ensemble, último acesso em 17/07/2023

https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics, último acesso em 17/07/2023