ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

Alexandre Castelo Branco Félix de Andrade

Iago Jordão Lipovetsky

José Milton de Andrade Rios Neto

MISSING VALUES: ESTIMATIVA DE VIAGENS MUNICIPAIS USANDO DADOS DE SISTEMAS DE IDENTIFICAÇÃO AUTOMÁTICA DE VEÍCULOS E APRENDIZADO DE MÁQUINA

Trabalho de formatura

do curso de Engenharia Civil

da Escola Politécnica da Universidade de São Paulo

Orientador: Prof. Dr. Cláudio Luiz Marte

São Paulo - 2020

RESUMO

A matriz origem-destino é essencial para o planejamento estratégico dos transportes e suas operações. Métodos tradicionais de estimação da matriz OD são, em geral, onerosos e demorados, dependendo da realização de diversas pesquisas de campo. O objetivo desse trabalho de formatura, em conjunto com o trabalho de mestrado de Douglas Martins, é propor uma metodologia de estimação da matriz OD utilizando dados de sistemas de identificação automática de veículos. O uso da infraestrutura pré-existente torna a aplicação da metodologia mais eficiente e menos onerosa.

Enquanto a dissertação de Douglas Martins se encarrega de diversas etapas da estimação da matriz OD, esse trabalho de formatura busca complementá-lo, estimando os valores faltantes ("*missing values*") da matriz obtida por ele. Por inúmeros fatores, como: erros de leitura das placas, falta de granularidade dos equipamentos e até erros da base de dados, em alguns casos, a metodologia de Martins não consegue obter nenhum valor para o fluxo entre pares OD. É justamente esse o objetivo desse trabalho de formatura, propor um método que visa estimar esses fluxos faltantes, por meio de técnicas de aprendizado de máquina ("*machine learning*") e, para tal, são utilizados dados socioeconômicos das zonas de origem e destino. O modelo é treinado usando os fluxos conhecidos e extrapolando são estimados os fluxos desconhecidos, obtendo uma matriz OD completa.

Vale destacar aqui que a ideia do trabalho não é obter uma matriz OD completa para São Paulo, mas sim obter um modelo que possa ser aplicado para diferentes regiões do Brasil e do mundo, utilizando diferentes bases de dados e diferentes dados para treinamento, dependendo do que estiver disponível para a região em estudo.

ABSTRACT

The origin-destination matrix is essential for the strategic planning of transport and its operations. Traditional OD matrix estimation methods are generally costly and time-consuming, relying on a series of field surveys. The bottom line of this graduation work, together with the master's qualification work of Mr. Douglas Martins is to describe and demonstrate the application of an estimation methodology for the OD matrix using data from automatic vehicle identification systems. The use of pre-existing infrastructure makes the application of the methodology more efficient and less costly

While Martins' work takes care of several stages of the OD matrix estimation, this graduation work seeks to complement it by estimating the missing values from the matrix obtained by Martins. Due to a number of reasons, such as: license plate recognition error, poor granularity and distribution of the equipments and even database errors, Martins' method cannot obtain any value for some OD pairs, that is exactly the goal of this graduation work. This study seeks to estimate the missing values through machine learning techniques and using demographic and economic indicators from the origin and destination zones. The model is calibrated using the known values and then it is used to extrapolate and estimate the unknown values or missing values, obtaining a complete OD matrix.

It is worth noting that the main goal is not to obtain a complete OD matrix for São Paulo, but to obtain a model that can be used in other regions of Brazil or the world, using different databases and different calibration data, depending on what is available for the study area. The idea is to create some sort of plug and play model for missing values estimation.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

Os estudos acerca da geração e atração de viagens se mostram como essenciais para o bom planejamento de uma cidade, auxiliando tanto na engenharia de tráfego quanto na tomada de decisão de investimentos em infraestrutura, não só relacionada a transportes, mas a equipamentos públicos, tais como escolas, hospitais, bibliotecas, entre outros. Nesse contexto, a cidade de São Paulo realiza decenalmente uma pesquisa Origem-Destino (OD), a fim estudar e entender a dinâmica de mobilidade dentro desta metrópole. Tal pesquisa é bastante onerosa, e depende de várias entrevistas e visitas de campo, o que a torna muito difícil de ser replicada em um intervalo de tempo menor.

Paralelamente a isso, percebe-se um aumento da utilização de Sistemas Inteligentes de Transportes (ITS), que incluem a telemática e todos os tipos de comunicação com veículos, entre veículos (por exemplo, carro-carro), e entre veículos e locais fixos, na operação e gestão da mobilidade urbana e interurbana. Inúmeras ferramentas estão hoje disponíveis para diferentes contextos e escalas, com aplicações que vão desde prioridade semafórica e sistemas de bilhetagem até gerenciamento do trânsito, abrangendo controle dos fluxos de carros, gerenciamento de incidentes relacionados às redes de transportes, gerenciamento da demanda, gerenciamento da manutenção da infraestrutura, fiscalização do cumprimento das regras de trânsito, dentre outros.

Tendo em vista as características que dificultam a realização das pesquisas OD e o crescente avanço tecnológico empregado no planejamento de transportes, diversos estudos buscam implementar métodos que proporcionem a geração dessas matrizes, sem depender de entrevistas ou visitas, usando a infraestrutura já existente, por meio de sistemas de identificação automática de veículos.

Utilizada em grande parte das aplicações de ITS, a identificação automática de veículos pode ser entendida como um conjunto de recursos de hardware, software e telecomunicações que interagem para atingir, do ponto de vista funcional, o objetivo de conseguir extrair digitalmente a identidade de um veículo. É feita tanto por sistemas que transmitem e recebem uma identidade visual quanto por sistemas que, instalados na

infraestrutura da via, são capazes de reconhecer a placa dos veículos circulantes (BERNARDI, 2015).

Todavia, a geração de matrizes OD a partir desses sistemas automáticos de identificação de veículos pode trazer células vazias, ou seja, sem um número de viagens entre um par Origem-Destino. Isso ocorre por diversos motivos, como falha nos equipamentos de identificação ou a pouca granularidade dos pontos de instalação desses equipamentos.

Este trabalho de formatura, portanto, visa a propor um método que faz uso de aprendizado de máquina para criar algoritmos que permitam estimar os valores faltantes para uma matriz OD. Ele foi aplicado à matriz OD da cidade de São Paulo, elaborada a partir de dados desses sistemas automáticos de identificação de veículos, e utiliza as características socioeconômicas das zonas OD da cidade para estimar os valores faltantes.

1.1. Contexto e Suporte

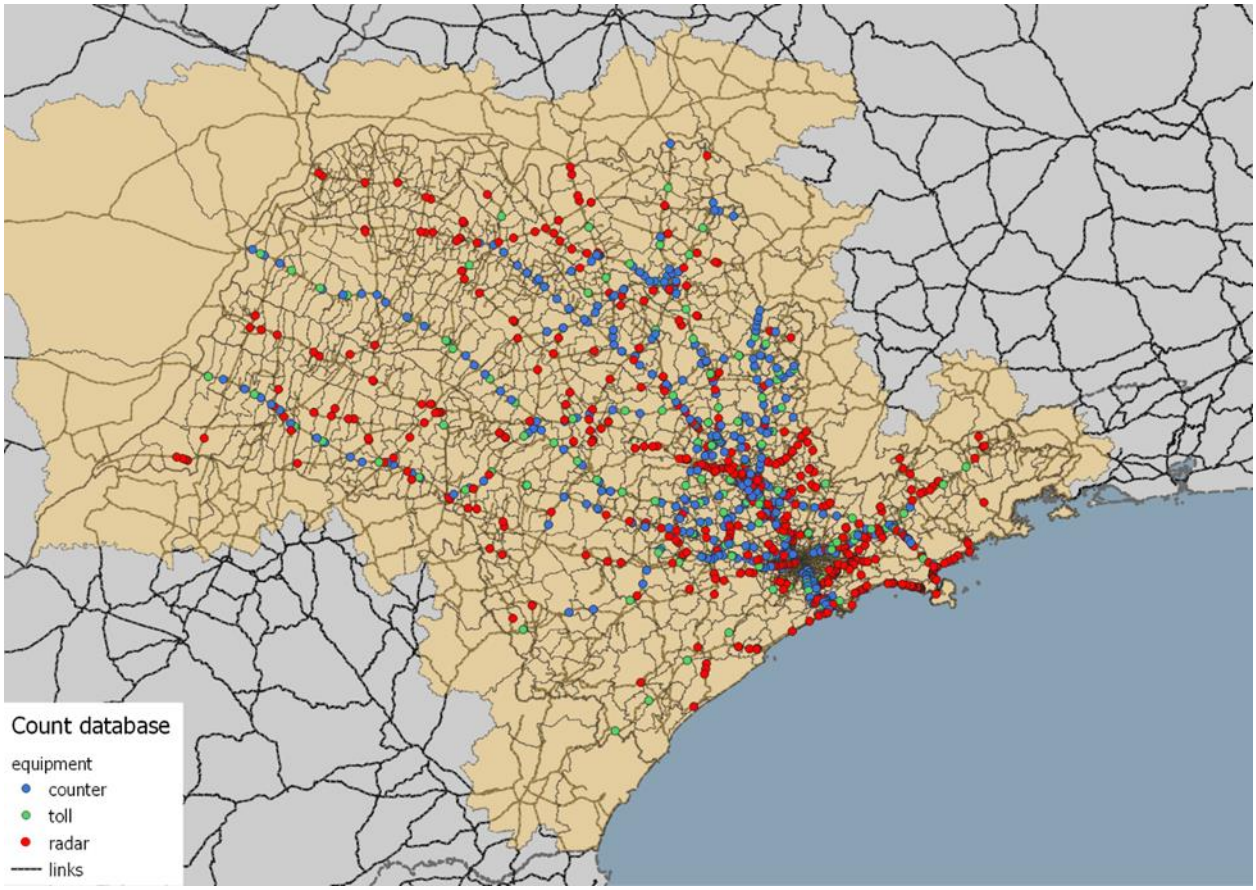
O trabalho de formatura foi desenvolvido utilizando como suporte o texto de qualificação da dissertação de mestrado: MARTINS, DOUGLAS F W CAPELOSSI. Scalable method for origin-destination demand estimation using automatic vehicle identification data (Método escalável de estimação da matriz origem-destino usando informações de identificação automática de veículos, em tradução livre). Esse trabalho de formatura busca complementar o estudo de Martins e para tanto faz-se necessária uma breve explicação acerca do que já foi desenvolvido pelo autor da dissertação, de modo que se possa compreender o contexto e a contribuição deste trabalho de formatura.

Martins tinha como principal objetivo descrever em detalhes e demonstrar a aplicação de um método para estimar a matriz OD baseado no reconhecimento das placas dos veículos, utilizando radares de trânsito e passagem dos veículos em pista automáticas de pedágio. A metodologia proposta por Martins ainda está em construção,

mas é eficiente, uma vez que se utiliza da infraestrutura pré-existente de radares de fiscalização e informações de passagens de veículos em pistas de pedágio.

O teste da metodologia foi realizado por Martins em todo o Estado de São Paulo, por meio de uma rede que compreende grande parte do tráfego do estado. A área de estudo conta com mais de 2 mil pontos de coleta de dados e mais de 6 bilhões de registros de veículos em 2018 – ver Figura 1.

Figura 1 - Pontos de coleta de dados disponíveis



Fonte: Martins (2019)

O método de Martins foi dividido em quatro principais etapas, cada uma delas gerando resultados essenciais para o próximo passo – ver Figura 2.

Figura 2 - Estrutura do método de Martins

Fonte: Martins (2019)

A primeira etapa consiste em transformar a base de dados de identificação automática de veículos em sub-rotas, por meio da identificação sequencial dos equipamentos (radares) que reconhecem determinado veículo. A capacidade de reconhecimento do sistema de radares depende de diversos fatores e pode resultar em leituras erradas ou até a não leitura de determinada placa.

Para melhorar a confiabilidade dos dados, Martins incorporou em seu estudo dados de passagem em pistas automáticas das praças de pedágio, que embora tenham uma granularidade bem reduzida, apresentam dados com acurácia muito maior. Idealmente, ter uma rede de identificação com equipamentos em cada esquina seria desejável e poderia resultar em uma matriz muito mais fidedigna, no entanto, o modelo de Martins parece promissor mesmo em regiões em que a granularidade do sistema está longe da desejável.

Um passo intermediário ainda da primeira etapa envolve o reconhecimento de erros e eventuais correções. Caso um veículo tenha percorrido uma rota em menos tempo do que se poderia fazer no melhor dos casos, provavelmente houve algum erro durante a identificação desse veículo. Erros como esse foram identificados e retirados do estudo por Martins.

Em seguida, na segunda etapa, um conjunto de possíveis pares OD é selecionado para cada uma dessas sub-rotas da etapa 1. Para tanto, critérios como o primeiro e o último equipamento a reconhecer o veículo, bem como caminhos prováveis entre pares OD foram considerados. Desse modo, caso exista outro caminho muito melhor entre um par OD, dificilmente algum indivíduo utilizaria a sub-rota analisada e, portanto, esse par OD não deveria estar entre os prováveis para a sub-rota em análise. Como a segunda

etapa apenas elege os pares OD mais prováveis, faz-se necessário um modelo para distribuir o total de viagens dessa sub-rota em cada par OD (terceira etapa).

A terceira etapa consiste em, utilizando um modelo gravitacional, determinar para o total de viagens em cada sub-rota a parcela pertencente a cada par OD. O modelo gravitacional em questão usa indicadores demográficos e econômicos para estimar a atração de cada zona.

Na quarta e última etapa do estudo de Martins, a matriz OD obtida é calibrada usando o algoritmo T-Flow Fuzzy do software VISUM. O algoritmo usa dados da contagem de veículos e corrige o volume total de viagens entre cada par OD, de modo que a matriz represente os volumes corretamente.

Por fim, o produto da dissertação de Martins é uma matriz OD gerada de forma quase automática, por meio dos dados provindos dos equipamentos de identificação automática de veículos. Método que pode ser facilmente escalável e aplicável em outras regiões a depender da disponibilidade de equipamentos de identificação de veículos.

Vale destacar aqui que, assim como qualquer metodologia, o estudo de Martins enfrenta alguns problemas, principalmente relacionados à imprecisão dos equipamentos de fiscalização (radares), bases de dados errôneas e regiões com baixa disponibilidade de equipamentos. Em alguns dos casos, a falta de dados é tamanha que o modelo de Martins não consegue obter nenhum volume de tráfego entre determinados pares OD, são os "*missing values*". Tentar contribuir para mitigar esse problema é justamente o objetivo deste trabalho de formatura.

Busca-se neste trabalho, estimar os chamados "*missing values"* por meio do aprendizado de máquina ("*machine learning"*), sendo que o método se utiliza dos próprios volumes obtidos satisfatoriamente por Martins para treinamento. Além dos dados de Martins, buscando ter uma maior precisão na estimação dos "*missing values*" o grupo utilizou também de dados socioeconômicos das zonas OD obtidos na pesquisa OD 2017 do Metrô de São Paulo.

O grupo acredita que a metodologia clássica (modelo de quatro etapas) ainda seja essencial e adotada pelas principais empresas de transporte, agências governamentais e profissionais. Mas, ressalta aqui que o método proposto por Martins foi validado por especialistas e tem obtido resultados promissores relacionados a sua aplicabilidade, tendo o diferencial de ser muito mais eficiente na coleta de dados.

1.2. Justificativa

A matriz origem-destino é essencial para o planejamento estratégico dos transportes e suas operações, bem como para o planejamento da cidade de maneira geral. Métodos tradicionais de estimação da matriz OD são, em geral, onerosos e demorados, dependendo da realização de diversas pesquisas de campo. O uso da infraestrutura pré-existente torna a aplicação do método proposto neste trabalho de formatura mais eficiente e menos custoso, podendo-se obter com uma frequência maior a estimação dos fluxos entre pares OD, o que pode ser útil para os órgãos responsáveis pela coordenação e pelo planejamento do trânsito na cidade.

A cidade de São Paulo, com suas características urbanísticas de modelo radial, tem um grande desafio no que tange à mobilidade urbana, já que há uma grande densidade de postos de trabalho e equipamentos de uso público nas regiões centrais, enquanto a maior parte da população que ocupa esses postos e faz uso desses equipamentos habita em áreas mais periféricas ou, até, em outros municípios. Tal contexto, faz com que, diariamente, milhões de pessoas necessitem cruzar a cidade ou entrar nela, demandando uma infraestrutura de transporte robusta e que consiga comportar e alocar tal movimentação de maneira eficiente.

De fato, a mobilidade nas grandes metrópoles do mundo é um desafio devido aos altos fluxos e a alta demanda. Além disso, essas cidades sofrem rápidas transformações em suas dinâmicas socioeconômicas, novos centros de geração e de atração de viagens surgem a cada ano. Assim, um intervalo muito amplo para se ter uma fotografia de onde a população da cidade sai e para onde ela vai pode ser muito danoso, não só para o

planejamento urbano, mas também para a iniciativa privada. Isso mostra que uma alternativa que sirva para estimar a origem e o destino da população, sem o ônus de realizar uma pesquisa completa, pode ser muito vantajosa e importante.

1.3. Objetivos

Este trabalho de formatura visa elaborar um método que faz uso de aprendizado de máquina para estimar os valores faltantes de uma matriz Origem-Destino para a Cidade de São Paulo. Este método é elaborado a partir dos registros de radares de tráfego no período de um ano (2018). Para isso, além dos fluxos observados, o método fará o uso de parâmetros, como o custo da viagem, o qual se baseia tanto no tempo quanto na distância percorrida, junto às características socioeconômicas de cada zona, disponibilizadas pela Pesquisa OD de 2017 para a Cidade de São Paulo, visto que acredita-se na existência de influência de tais aspectos sobre a dinâmica de mobilidade urbana de uma da região.

Com isso, pretende-se criar um método que possa ser escalável e replicável em diversas cidades do país e permita, de maneira menos custosa, expor a dinâmica de transporte e de mobilidade da área em que se está implementando, o qual pode ser desenvolvido e chegar a auxiliar órgãos e secretarias de tráfego a realizar um melhor planejamento da infraestrutura de transporte da região.

2. DADOS UTILIZADOS

O método desenvolvido por esse trabalho de formatura, como qualquer método de aprendizado de máquina, depende fortemente dos dados disponíveis para calibração e treinamento do modelo. Caso a qualidade dos dados não seja satisfatória, o modelo criado também não será promissor.

O grupo se utilizou de duas principais fontes de dados: a Secretaria de Logística e Transporte do Estado de São Paulo e a Pesquisa OD do Metrô de São Paulo.

Da Secretaria de Logística e Transporte do Estado de São Paulo foram utilizados dados de volume (quantidade de veículos viajando entre pares OD) e custo da viagem entre o par OD. Vale destacar aqui que são justamente os dados de volume, agora mencionados, que contém os valores faltantes entre alguns pares OD e que este trabalho busca estimar.

Da pesquisa OD do Metrô de São Paulo foram utilizados dados socioeconômicos. Após o treinamento do modelo, é a partir desses dados socioeconômicos que os volumes faltantes são estimados.

2.1. Dados Agregados de Identificação Automática de Veículos

Os dados recebidos da Secretaria de Logística e Transporte do Estado de São Paulo incluem uma matriz OD gerada por meio de dados de sistemas automáticos de identificação de veículos (disponível no Google Drive criado pelo grupo e cujo link se encontra no item 7 desse trabalho). A matriz abaixo (Figura 3) tem propósito apenas ilustrativo do ponto de partida do grupo.

Figura 3 – Matriz OD ilustrativa

O\D	A	B	C	D	E
A	6	7	4	**?**	8
B	2	**?**	4	8	**?**
C	2	6	2	10	5
D	9	**?**	7	3	2
E	6	5	2	**?**	8

Fonte: Elaboração própria (2020)

A matriz recebida, conforme exemplificado acima, contém o fluxo diário de veículos, em milhares, entre uma origem e um destino (no caso, aproximadas por radares). A matriz original utilizada no trabalho tem 377 linhas e 377 colunas, totalizando mais de 140 mil pares OD, e, por isso, não foi mostrada aqui (embora esteja disponível no Google Drive criado pelo grupo, cujo link está disponível no item 7 deste trabalho). Os fluxos identificados pelos pontos de interrogação ilustram justamente os *missing values* e representam cerca de 20% da matriz.

2.1.1. Radares

A cidade de São Paulo começou a utilizar radares com identificação automática de placas no final dos anos 2000. Esse avanço permitiu que os radares deixassem de apenas fiscalizar o excesso de velocidade dos veículos e passassem a identificar, também, ocorrências de circulação de veículos em dias de rodízio; de invasões de faixas exclusivas de ônibus ou de zonas restritas à circulação de caminhões; de avanços de sinal vermelho; de paradas sobre a faixa de pedestres no fechamento de semáforos e, por fim, de conversões proibidas (BERNARDI, 2015). Mais tarde, com a associação

desses sistemas a bases de dados atualizadas, tornou-se possível, ainda, detectar a passagem de veículos com irregularidades administrativas.

Apesar desse avanço, os sistemas de identificação automática de veículos ainda não são utilizados amplamente para a coleta de dados de tráfego em tempo real. É possível coletar e transmitir dados de velocidade instantânea, volume de veículos, comprimento do veículo e ocupação, por exemplo, que podem auxiliar a gestão do tráfego em vias movimentadas (BERNARDI, 2015). Além disso, pode-se utilizar alguns desses dados para comunicar aos motoristas a situação de trânsito ou o tempo de percurso estimado até certo destino, permitindo que tomem decisões quanto a possíveis rotas alternativas. Conforme Bernardi (2015), isso só depende de sistemas de monitoramento adequados para fornecer, dinamicamente, as informações desejadas. Como mencionado anteriormente, uma possibilidade de uso desses sistemas de radares é a geração de dados para a construção de matrizes de origem e destino praticamente em tempo real, que, no entanto, ainda exige a solução de algumas questões.

2.1.2. Custo de Viagens

É um dos parâmetros mais importantes para a constituição do modelo. O custo da viagem representa uma estimativa dos recursos dispendidos para se realizar a viagem entre um par OD. Tal custo busca englobar não só o dispêndio financeiro empregado no percurso, mas também o tempo gasto para o percorrer, o qual muitas vezes é incerto, já que o método utilizado não leva em conta fatores como congestionamentos, bastante presentes na cidade e que sempre alteram o tempo da viagem. Nesse trabalho de formatura foi assumido que o custo pode ser aproximado pelo caminho de menor distância, em rede, entre o par OD dividido pela velocidade regulamentada das vias. Obtendo assim um custo aproximando pelo tempo da viagem.

A Tabela 1, disponibilizada pela Secretaria de Logística do Estado de São Paulo, conta com 3.878.930 amostras e possui 6 atributos, sendo:

- **fid (from-id)**: número de identificação da origem

- **tid (to-id)**: número de identificação da chegada
- **cost**: custo da viagem
- **pass_avi**: ”flag” de identificação de passagem por radar contendo identificação veicular automática
- **sent12**: sentido da viagem
- **cost_f**: custo da viagem

Podendo ser exemplificada como:

Tabela 1 – Exemplo de entradas da tabela *Cost_min*

fid	tid	cost	pass_avi	sent12	cost_f
31021	31022	2.115278	no	1	2.12
31021	31023	3.837778	no	1	3.84
31021	31047	4.004167	no	1	4.00
31021	31048	3.617223	no	1	3.62
31021	31051	4.883617	yes	1	4.88

Fonte: Elaboração própria (2020)

2.1.3. Volume de Viagens

Para se obter o volume de viagens entre zonas OD considerou-se uma aproximação de que um radar dentro de uma zona representa a própria zona, e utilizou-se das observações de um ano desses radares. Esses dados foram disponibilizados pela Secretaria de Logística do Estado de São Paulo, por meio da base de dados do DETECTA, que consolida todas as informações aferidas pelos radares e sensores de tráfego do Estado. Ao consolidar essas informações, no entanto, a secretaria não obteve nenhuma estimativa para o fluxo de cerca da 20% dos pares OD analisados e são justamente esses dados que se busca estimar neste trabalho de formatura.

Para o número de viagens foram disponibilizadas duas tabelas semelhantes, mas que se diferem ao agregar os veículos: uma agrega todos os veículos (Tabela 2 - *paths_allmun*) e a outra agrega, pelo menos, os 10 maiores municípios registrados para cada veículo (Tabela 3 - *paths_muntop10*).

Ambas possuem atributos que descrevem o primeiro ponto de identificação dos veículos – *first* – e o último – *last*. Além disso, possuem atributos “cat_N”, que representam o número de viagens para cada tipo de veículo, por meio do fluxo diário médio de um ano, em milhares de veículos. A variável relevante para nosso trabalho é o “cat1”, pois se restringe às viagens atribuídas apenas aos carros.

Tabela 2 - Exemplo de entradas da tabela *paths_allmun*

first	last	cat1	cat2	cat3	cat4	cat5	cat6	catTot
3.550308e+09	3.550308e+09	1126.325	54.383	49.572	10.290	21.352	20.650	1282.572
3.550308e+09	3.550308e+09	827.568	36.241	24.317	4.867	13.308	10.017	916.318
2.019700e+04	2.019700e+04	1214.532	51.694	26.361	4.999	12.905	9.414	1319.905
3.550308e+09	3.550308e+09	565.141	21.425	15.607	2.290	5.319	4.269	614.051
1.026200e+04	1.032800e+04	714.552	10.593	2.576	0.412	0.074	0.051	728.258

Fonte: Elaboração própria (2020)

Tabela 3 - Exemplo de entradas da tabela *paths_muntop10*

first	last	city_name	cat1	cat2	cat3	cat4	cat5	cat6	catTot	catTot_group
3550308019	3550308019	SAO PAULO	540.466	16.506	12.885	2.359	3.825	2.297	578.338	1282.572
3550308019	3550308019	GUARULHOS	167.290	4.940	3.115	0.760	1.926	0.902	178.933	1282.572
3550308019	3550308019	0	87.783	7.419	17.469	4.461	8.104	11.130	136.366	1282.572
3550308019	3550308019	SAO JOSE DOS CAMPOS	26.602	0.394	0.235	0.041	0.175	0.099	27.546	1282.572
3550308019	3550308019	OSASCO	17.550	1.593	1.292	0.060	0.648	0.259	21.402	1282.572

Fonte: Elaboração própria (2020)

Enquanto que a Tabela 2 - *paths_allmun* apresenta 711.941 entradas e o registro de um total de 451.762 viagens na “cat1”, a Tabela 3 - *paths_muntop10* apresenta 3.557.095 entradas, registrando um total de 406.894 viagens por carro (na “cat1”). No entanto, dessas viagens registradas na Tabela 3 - *paths_muntop10* apenas 3.191.979 são de veículos provenientes de dentro do estado, contabilizando dessa forma apenas 371.064 viagens de carro com veículos de outros estados do país.

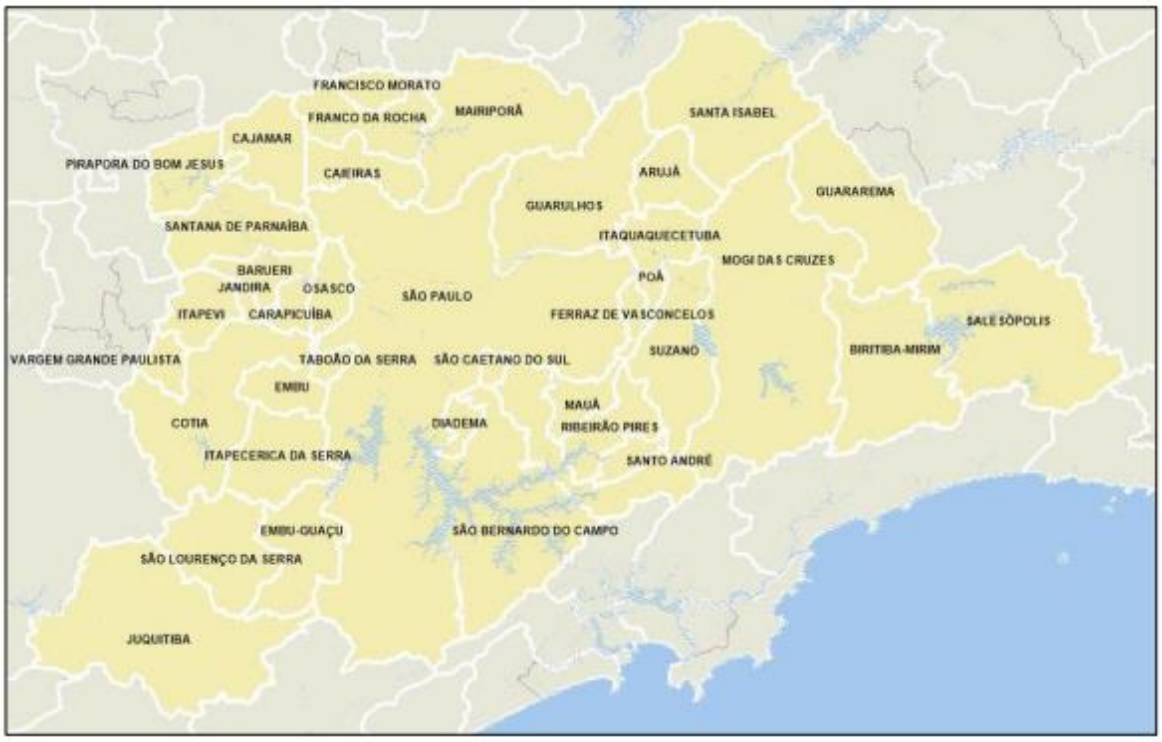
2.2. Dados Socioeconômicos

A Pesquisa Origem-Destino de 2017 do metrô de São Paulo foi utilizada como fonte para os dados socioeconômicos utilizados nesse trabalho de formatura (Figura 4). Os dados em questão foram fundamentais para calibrar as correlações e, em seguida, estimar os fluxos inexistentes.

Os principais dados utilizados são relativos a:

- População
- Renda
- Empregos
- Posse de Automóveis

Figura 4 – Área de abrangência da Pesquisa OD

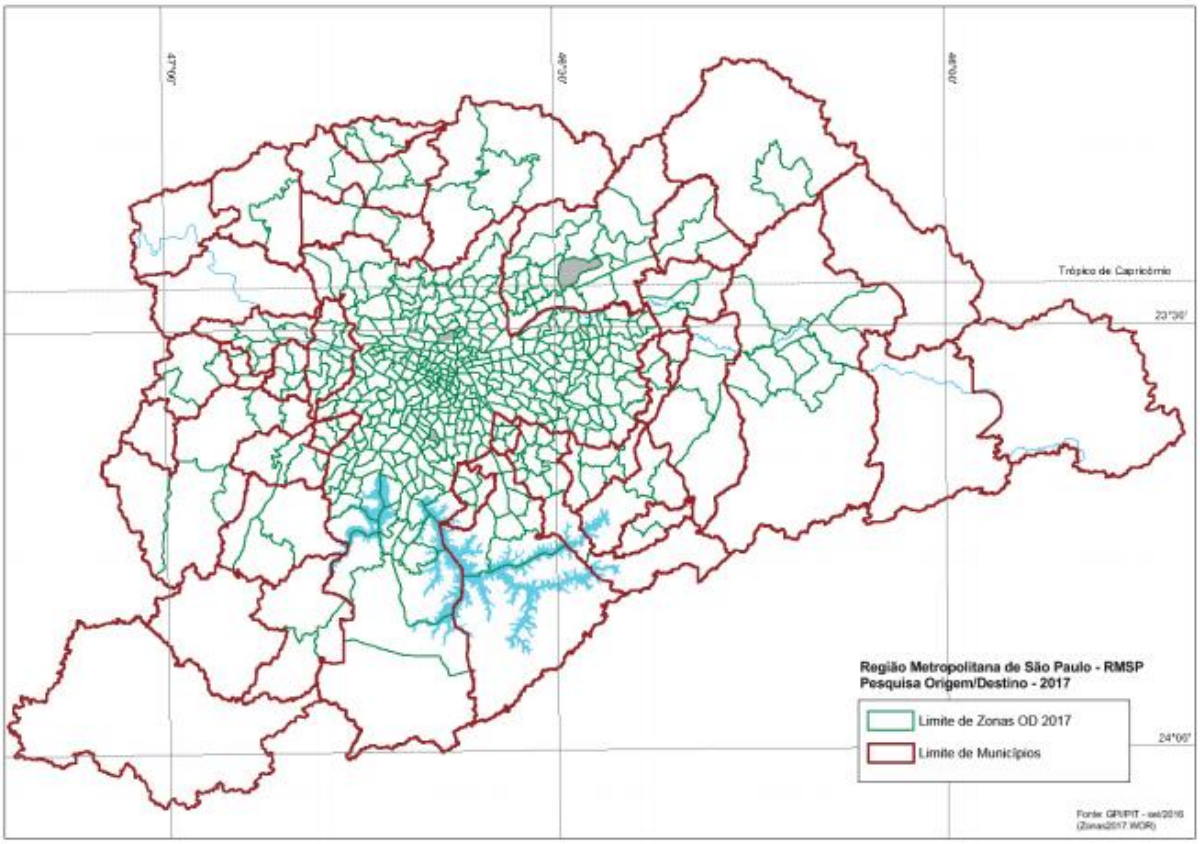


Fonte: Pesquisa OD (2017)

A pesquisa abrange toda a região metropolitana de São Paulo, no entanto, como a granularidade dos radares disponíveis diminui bastante ao se afastar da capital, o grupo optou por reduzir o escopo desse trabalho para incluir apenas o município de São Paulo.

A pesquisa divide a sua área de abrangência em 517 zonas OD, das quais 342 pertencem ao município de São Paulo e foram, portanto, utilizadas para esse trabalho.

Figura 5 – Zoneamento Pesquisa OD



Fonte: Pesquisa (2017)

Depois de divididas, as zonas são caracterizadas por diversos aspectos socioeconômicos. Lista-se no item 720, em anexo, as variáveis utilizadas nesse trabalho de formatura.

3. METODOLOGIA

3.1. Diferentes abordagens para a imputação de valores faltantes

- **Remoção dos valores faltantes**

Uma abordagem direta para a falta de dados é excluí-los. No contexto de uma futura regressão, geralmente significa a necessidade de uma análise de caso completo: excluir todas as unidades para as quais o resultado ou qualquer uma das entradas está faltando. Essa abordagem pode gerar dois grandes problemas:

  - A ordem de grandeza dos valores ausentes (valores práticos reais) e dos valores presentes sejam tão diferentes tal que possa enviesar a amostra.
  - Se há muitas variáveis, haverá poucos casos completos e, portanto, a maior parte dos dados poderá ser descartada.

- **Substituição dos valores faltantes**

Nesta abordagem opta-se por usar uma estimativa dos valores faltantes ao invés do descarte dos dados. Manter o tamanho completo da amostra é vantajoso quanto ao viés e à precisão, porém dependendo da estimativa esta vantagem pode ser perdida pois podem ser gerados diferentes tipos de viés.

Uma série de métodos podem ser utilizados, de modelos mais simples à mais complexos. Dentre alguns métodos, podem ser citados:

1. Média
2. Mediana
3. Valor Anterior
4. Informações de observações da base de dados
5. Modelos de aprendizado de máquina supervisionado
6. Redes Neurais

A imputação por média, mediana ou mesmo pelo valor anterior apenas examina a distribuição dos valores da variável com entradas ausentes. Se sabemos que existe uma correlação entre o valor ausente e outras variáveis, geralmente podemos obter melhores suposições, regredindo a variável ausente a partir de outras variáveis.

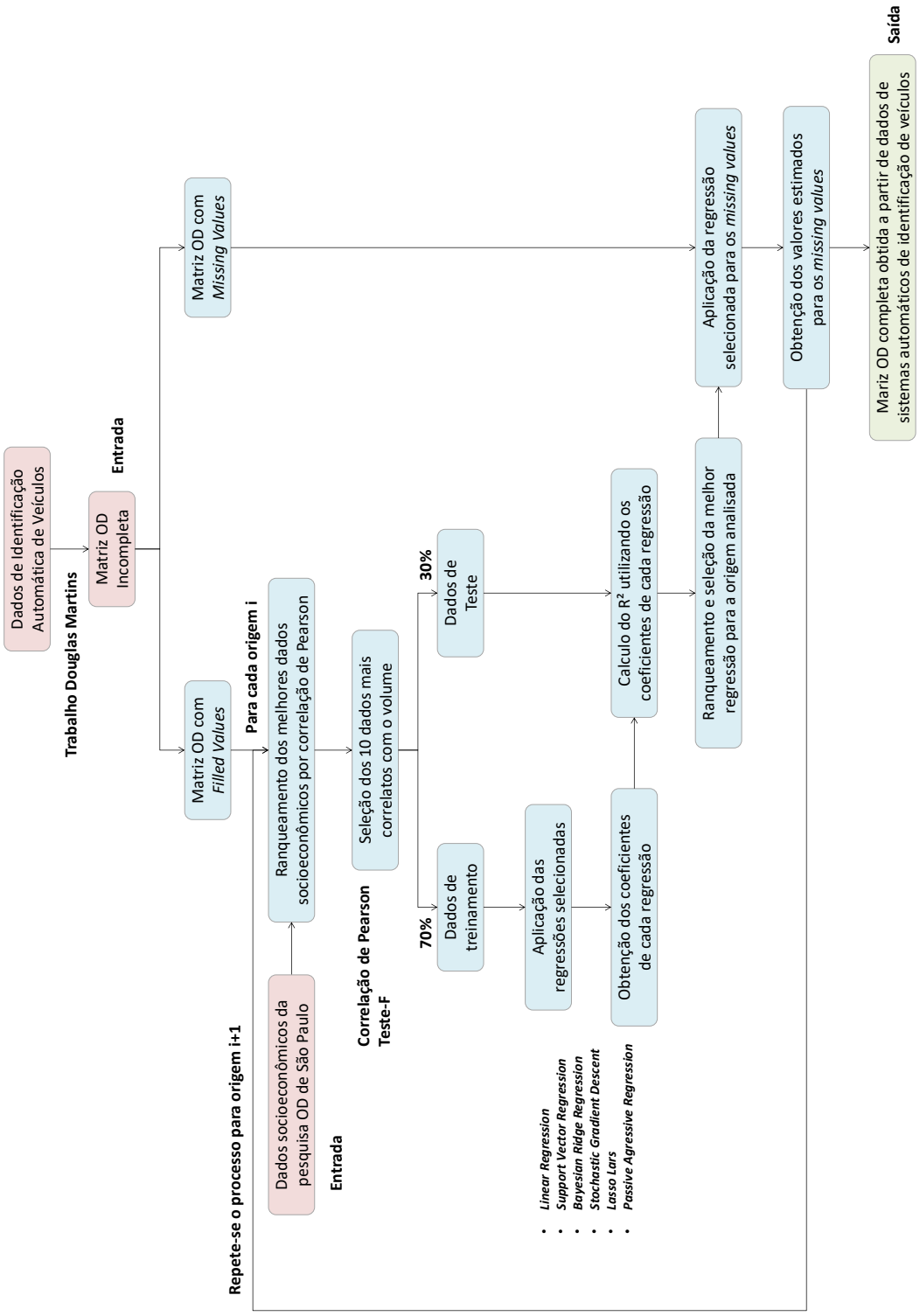
Com métodos de regressão, as informações de outras variáveis são usadas para prever os valores ausentes em uma variável usando um modelo como de Regressão Linear, por exemplo. Geralmente, primeiro o modelo de regressão é estimado nos dados observados (treinado) e, posteriormente, usando os pesos de regressão, os valores ausentes são previstos e substituídos.

Para este trabalho de formatura adotaremos, como explicaremos posteriormente, diferentes modelos de regressão em vez de um método único. Como uma abordagem geral, a ideia é montar uma estrutura para o procedimento de imputação, que será executado várias vezes, de forma a criar diferentes conjuntos de dados imputados plausíveis. A principal motivação para usar múltiplas imputações é que uma única regressão pode não refletir a variabilidade dos dados da amostra e dos valores ausentes.

3.2. Etapas

Encontra-se na Figura 6 um fluxograma simplificado da metodologia utilizada neste trabalho de formatura, sendo que cada um dos passos será descrito nos tópicos seguintes.

Figura 6 – Fluxograma da Metodologia



Fonte: Elaboração própria (2020)

- **Dados de Identificação Automática de Veículos**

Os dados utilizados foram descritos em detalhes na seção 2.1.

- **Matriz OD Incompleta**

A qualificação de mestrado de Martins processou os dados dos radares e obteve uma matriz OD. No entanto, como já explicado na seção 1.1, essa matriz obtida por Martins contempla alguns valores faltantes (“*missing values*”). É justamente aqui que começa o escopo desse trabalho de formatura.

- **Matriz OD com *Missing Values***

Corresponde a divisão da matriz que contem apenas os valores faltantes, essa subdivisão da matriz OD incompleta será utilizada na fase de estimação dos “*missing values*”.

- **Matriz OD com *Filled Values***

Corresponde a divisão da matriz OD incompleta que contem apenas os valores preenchidos da matriz. É essa subdivisão da matriz que será utilizada em sequência.

- **Eliminação de Dados Indesejados**

Foi utilizada a regra de Turkey (descrita na seção 3.4.1) para eliminação de dados indesejados. A partir dessa etapa, o processo se repete para cada origem “i” analisada. Ou seja, o modelo foi estratificado por origem.

- **Dados Socioeconômicos da Pesquisa OD de São Paulo**

Nesse momento, são inseridos no modelo os dados socioeconômicos descritos na seção 2.2.

- **Ranqueamento dos Dados Socioeconômicos por Correlação de Pearson**

Dentre os 106 dados socioeconômicos inseridos no modelo, utiliza-se a correlação de Pearson para entender quais são os mais correlacionados com o volume que se deseja estimar. Destaca-se aqui que, nesse momento, também são analisadas as correlações dos dados socioeconômicos entre si, de modo a evitar a utilização de dados altamente correlacionados. Esse processo está descrito em detalhes na seção 3.5.

- **Seleção dos 10 Dados Mais Correlacionados**

Nessa etapa, são selecionados os 10 dados que tiveram melhor correlação com o volume, os quais são submetidos ao Teste F, que avalia se a inclusão de uma variável é relevante ou dispensável para modelar uma dada origem. Somente esses 10 dados socioeconômicos serão utilizados a partir de agora. Todos os outros 96 dados serão descartados. Um aprimoramento desse trabalho de formatura poderá ser tornar esse procedimento adaptável a determinar o melhor conjunto de dados socioeconômicos necessários para estimar cada par O/D, não se restringindo a unicamente 10.

O Teste F está explicado, detalhadamente, no item 3.5 deste trabalho de formatura.

- **Dados de Teste**

Nesse momento, a matriz OD de *Filled Values* é subdividida, sendo que 30% dos valores serão destinados aos testes, isto é, esta parcela serve para avaliar como o

modelo se comporta, se estima com precisão os volume que, nesse caso, já são conhecidos.

- **Dados de Treinamento**

70% dos valores da matriz OD de *Filled Values* são utilizadas para treinar e calibrar o modelo. Essa divisão da matriz entre dados de teste e de treinamento é realizada aleatoriamente. Se algum dos dados de teste fosse utilizado para treinamento, seriam obtidos modelos enviesados.

- **Pré-processamento**

Esta etapa está descrita em detalhes na seção 3.6. Serve para padronizar os dados utilizados, de modo a evitar que apenas uma variável seja significativa para o modelo.

- **Escolha dos Hiperparâmetros**

Os hiperparâmetros são as variáveis que controlam o próprio processo de treinamento do modelo. Esta etapa detalhada na seção 3.8

- **Aplicação das Regressões Selecionadas**

As regressões utilizadas nesse trabalho estão descritas em detalhes na seção 3.7. Nessa etapa, são aplicadas as regressões utilizando o volume ou fluxo entre pares OD e os dados socioeconômicos de cada zona.

- **Obtenção dos Coeficientes de cada Regressão**

Aqui são obtidos os coeficientes de cada regressão, esses coeficientes serão posteriormente utilizados para estimar os valores dos dados de teste e entender como cada regressor se comporta.

- **Eliminação das Alternativas cuja Curva de Aprendizagem Não Converge**

Etapa detalhada na seção 3.9.

- **Cálculo do R^2 utilizando os Coeficientes Obtidos para cada Regressão**

Utilizando os coeficientes já obtidos de cada regressão, são calculados os valores de volume ou fluxo dos dados de teste. Os valores calculados são comparados com os valores observados, de modo a entender se o modelo se comporta bem, ou seja, consegue estimar com precisão os volumes.

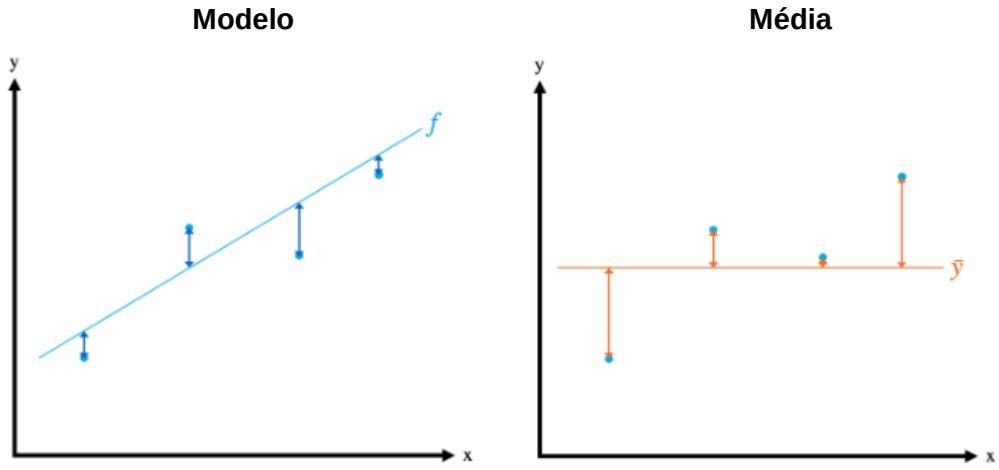
Para comparar as alternativas de regressão entre si, precisa-se utilizar uma métrica. Nesse trabalho, optou-se por utilizar o R^2 (Equação 1), que mede o quanto a predição do modelo é boa em comparação com a média.

- Quanto mais próximo de 1, melhor a predição.
- Se < 0, o modelo pode ser considerado inútil, pois seria mais próximo da realidade ter usado a média para estimar os valores faltantes ("*missing values*"). Ver Figura 7

Equação 1

$$R^2 = 1 - \frac{\mathrm{MSE(model)}}{\mathrm{MSE(baseline)}}$$

Figura 7 – Exemplo de Aplicação da Métrica R^2



Fonte: MQL5 (2017)

- **Ranqueamento e Seleção da Melhor Regressão para a Origem Analisada**

A partir da comparação entre os R^2 de cada regressão, seleciona-se aquela que obteve maior pontuação. Essa será a regressão de fato utilizada para estimar os valores faltantes (“*missing values*”)

Esse processo, desde a etapa de “ranqueamento dos dados socioeconômicos por correlação de Pearson” se repete para cada uma das origens analisadas.

- **Aplicação da Regressão Selecionada para os *Missing Values***

Nessa etapa, aplica-se a regressão que melhor se comportou utilizando os 10 dados socioeconômicos selecionados anteriormente para os valores faltantes (“*missing values*”)

- **Obtenção dos Valores Estimados para os *Missing Values***

Ao aplicar a regressão, obtém-se uma estimativa de volume ou fluxo dos pares OD faltantes.

- **Matriz OD Completa Obtida a partir de Dados de Sistemas Automáticos de Identificação de Veículos**

Finalmente, ao juntar a matriz de "*filled values"* com os valores estimados dos "*missing values*", o resultado é uma matriz OD completa, saída do modelo.

3.3. Preparação dos Dados

Para uma amostra de tamanho *n*, sendo X uma matriz *n* x *p*, contendo todas as variáveis sem valores faltantes, que serão utilizadas para predição destes valores faltantes. Esta amostra X consiste apenas de dados numéricos contínuos. Ver Tabela 4

Tabela 4 – Exemplo de tabela com os atributos utilizados no trabalho

fid	lid	cost_f	domicilios	familias	populacao	matriculas	empregos	automoveis	viagens_produzidas	...
12716	10006	0.72	7996.0	7996.0	18401.0	4010.2	12141.9	2582.2	40630.2	...
12716	11172	0.80	6836.0	6836.0	14277.0	1979.1	12564.2	7612.0	42750.3	...
12716	10836	0.69	12588.0	12588.0	24089.0	1126.7	22553.9	3400.9	78184.0	...
12716	10013	0.64	18909.0	19000.6	43094.0	11004.1	33133.3	11358.7	112648.0	...
12716	11064	0.65	1660.0	1660.0	3993.0	4180.4	19561.3	1300.1	53822.8	...

Fonte: Elaboração própria (2020)

Y é uma matriz *n* x *1* (Tabela 5) que consiste nos valores de volumes de tráfegos aos quais serão utilizados em uma matriz OD entre radares na cidade de São Paulo. Esta matriz contém dados numéricos contínuos. Nesta matriz, há ausência de valores, os quais suas estimativas são objeto de estudo deste trabalho de formatura.

Tabela 5 – Tabela de volumes entre pares OD

vol
1.933
0.500
0.333
0.067
1.800

Fonte: Elaboração própria (2020)

O nosso *dataset* se trata então, da combinação das variáveis X de entrada ou input e do valor de saída Y ou output.

Para aplicação dos modelos de regressão precisamos dividir este *dataset* em dois tipos de conjunto de dados, um para criar o modelo e outro para testar o desempenho do nosso modelo. O desempenho do nosso conjunto de testes não deve ser diferente do nosso conjunto de treinamento, o que significa que nosso modelo de aprendizado de máquina está se saindo bem e generaliza os exemplos de nosso conjunto de dados, em vez de aprendê-los em rotina (decorá-los).

Os tipos de conjunto de dados:

- **Conjunto de treinamento**

O conjunto de dados que usamos para treinar o modelo. O modelo vê e aprende com esses dados. Para este trabalho consideramos como 70% da amostra.

- **Conjunto de Teste**

O conjunto de dados de teste fornece o padrão-ouro usado para avaliar o modelo. Só é usado quando um modelo é completamente treinado (usando os conjuntos de treino e validação). Para este trabalho consideramos como 30% da amostra.

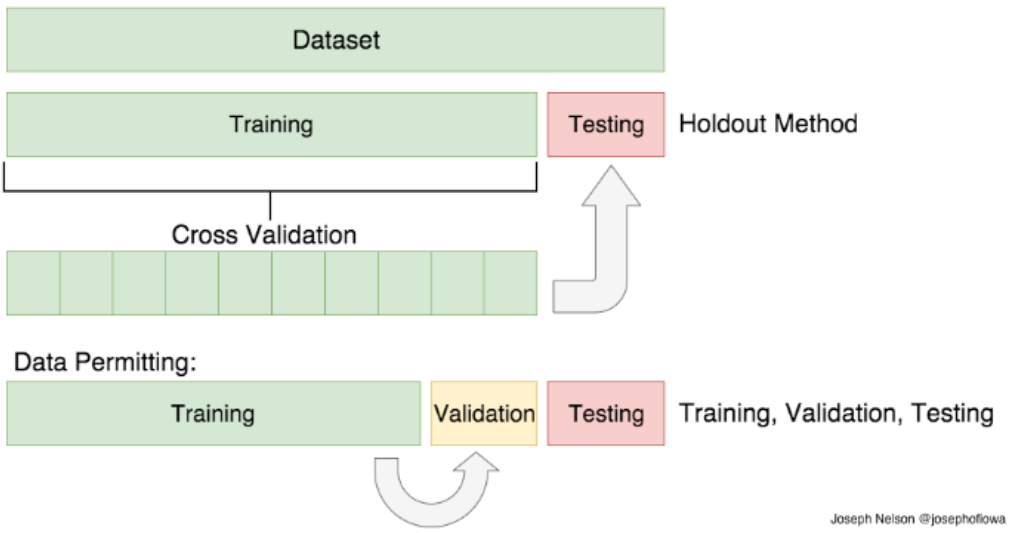
Devido haver a necessidade de execução de validações durante a seleção do modelo, seja para otimizar os hiperparâmetos, seja para verificação da capacidade de generalização do modelo, e da impossibilidade do uso do conjunto de teste sobre o risco de enviesar o modelo, usamos um conjunto de dados dentro do conjunto de treino que chamaremos de conjunto de validação.

- **Conjunto de Validação**

É a amostra de dados usada para fornecer uma avaliação imparcial de um modelo ajustado no conjunto de dados de treinamento ao se ajustar os hiperparâmetros do modelo. Para este trabalho de formatura foi adotado que conjunto de validação será 10% (*10 fold cross-validation*) da amostra de treinamento, sendo este conjunto escolhido aleatoriamente.

A avaliação se torna mais tendenciosa à medida que a habilidade no conjunto de dados de validação é incorporada na configuração do modelo. Ver Figura 8

Figura 8 – Divisão dos dados



Fonte: NELSON. J. (2016)

3.4. Pré-Processamento: Remoção de *Outliers*

Os *outliers* são valores extremos que se desviam de outras observações sobre os dados, podendo indicar uma variabilidade em uma medição, erros experimentais ou uma novidade. Em outras palavras, um *outlier* é uma observação que diverge de um padrão geral em uma amostra.

No processo de produção, coleta, processamento e análise de dados, os valores discrepantes podem vir de muitas fontes e se esconder em várias dimensões. Para remoção de *outliers* adotaremos como critério a regra de Tukey, explicada logo em seguida.

3.4.1. Regra de Tukey ou Intervalo Interquartil

Para entendermos melhor este método de remoção de dados indesejados precisaremos descrever cinco números importantes para seleção destes dados:

- *O valor mínimo ou mais baixo do conjunto de dados*
- *O primeiro quartil Q1 - representa um quarto de todos os dados*
- *A mediana do conjunto de dados - representa o ponto médio da lista de todos os dados*
- *O terceiro quartil Q3 - representa três quartos de todos os dados*
- *O valor máximo ou mais alto do conjunto de dados.*

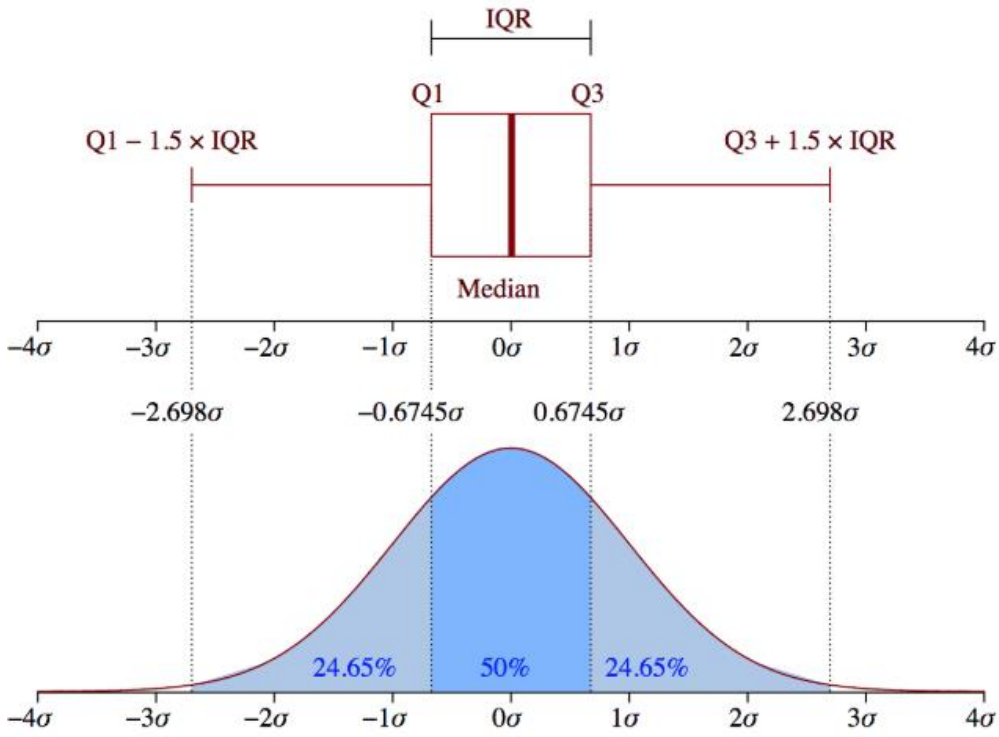
Esses cinco números podem ser usados para nos informar um pouco sobre nossos dados. Por exemplo, o intervalo interquartil é calculado subtraindo o primeiro quartil do terceiro quartil, servindo como indicador de como distribuir o conjunto de dados. (Equação 2)

Equação 2

$$IQR = Q3 - Q1$$

IQR é um conceito em estatística usado para medir a dispersão estatística e a variabilidade dos dados, dividindo o conjunto de dados em quartis. Em palavras simples, qualquer conjunto de dados ou qualquer conjunto de observações é dividido em quatro intervalos, definidos com base nos valores dos dados e em como eles se comparam ao conjunto de dados inteiro. O intervalo interquartil mostra como os dados são espalhados sobre a média. É menos suscetível a presença de valores indesejáveis. Ver Figura 9

Figura 9 – Regra de Tukey

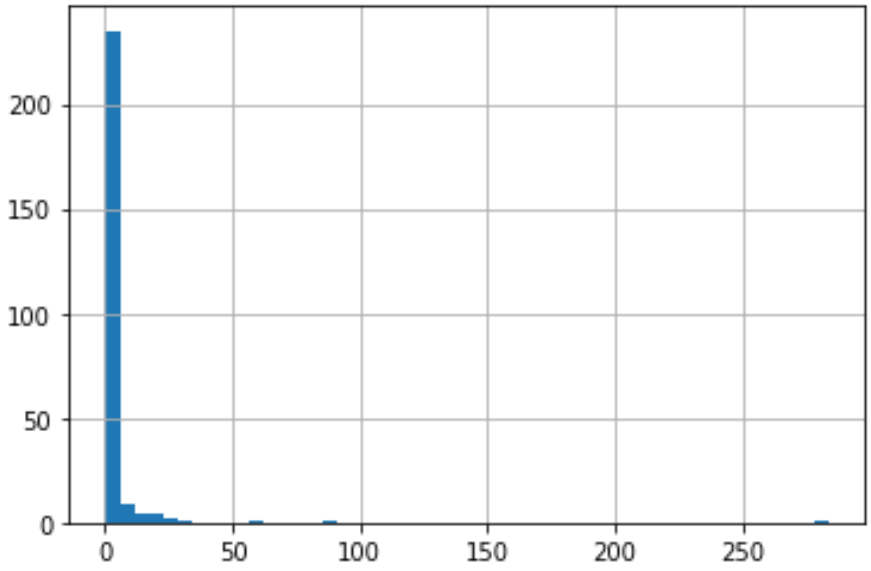


Fonte: NeuroMat (2013)

Os valores discrepantes neste caso são definidos como as observações que estão abaixo *(Q1 - 1,5x IQR)* ou traço inferior do *boxplot ,* ou estão acima de *(Q3 + 1,5x IQR)*, traço superior do *boxplot*.

Podemos observar na Figura 10 que os dados se concentram em valores muito baixos, numa distribuição inclinada à esquerda:

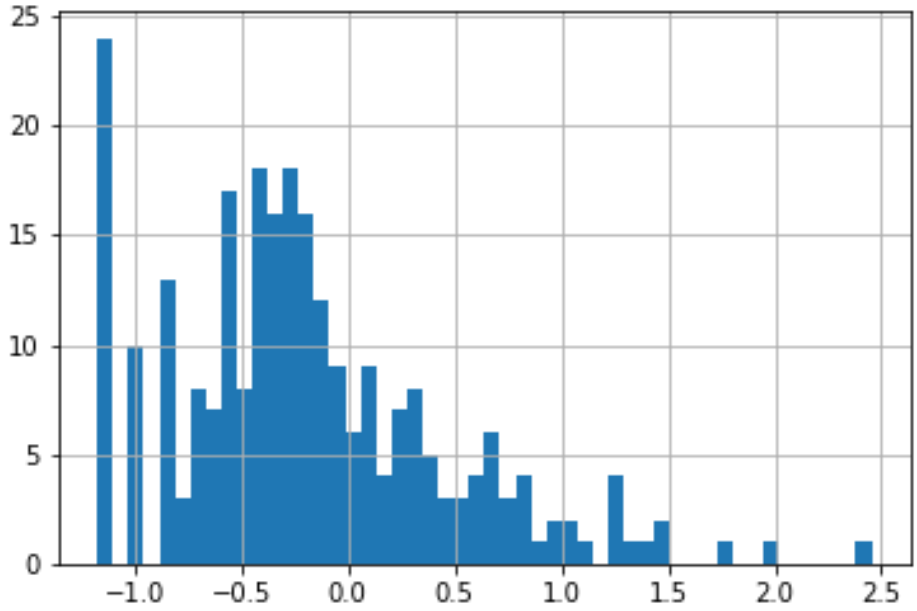
Figura 10 – Distribuição do *dataset selecionado* deste trabalho



Fonte: Elaboração própria (2020)

Uma transformação logarítmica (Figura 11) pode ajudar a ajustar uma distribuição muito distorcida em uma distribuição gaussiana. Após a transformação logarítmica, podemos ver facilmente a mudança de padrão em nossos dados.

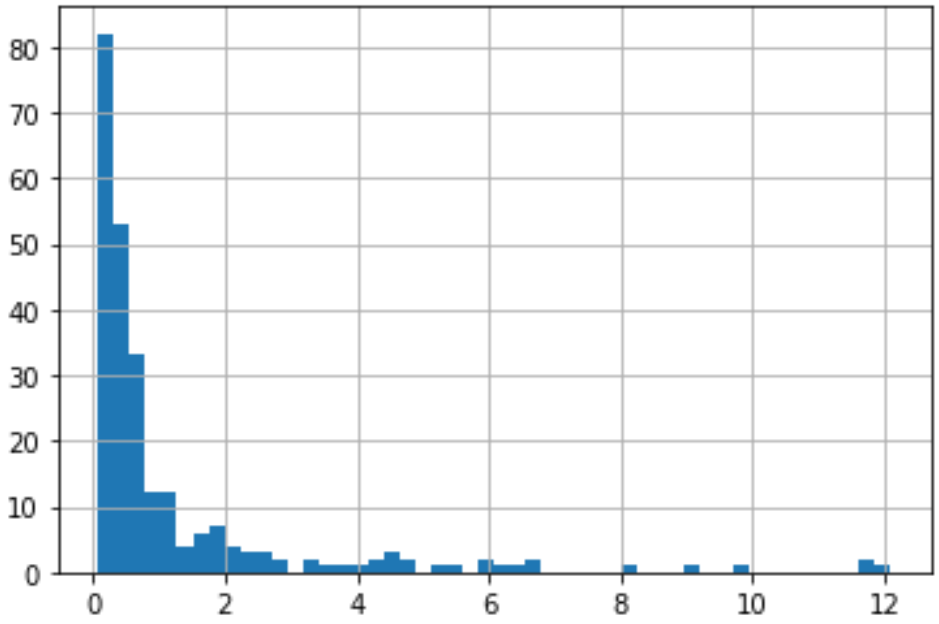
Figura 11 – Distribuição após transformação



Fonte: Elaboração própria (2020)

Com está distribuição aplicamos a regra de Tukey e retornamos os dados aos valores originais, eliminando assim os *outliers*. Ver Figura 12

Figura 12 – Distribuição após aplicação da regra de Tukey



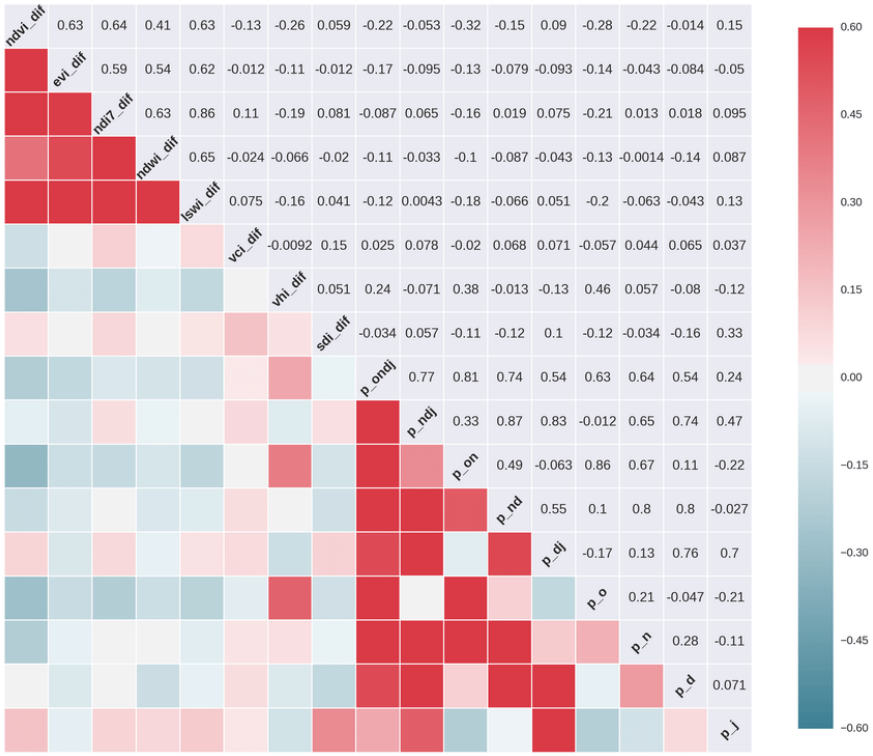
Fonte: Elaboração própria (2020)

3.5. Seleção dos dados mais correlacionados

A seleção de *features* (variáveis de entrada do modelo) é uma maneira de reduzir o número de variáveis e, portanto, reduzir a complexidade computacional do modelo. Muitas vezes, a seleção de *features* se torna útil para superar o problema de sobreajuste (*overfitting*). Isso nos ajuda a determinar o menor conjunto de variáveis necessárias para prever a variável de resposta com alta precisão.

Este é um problema de modelagem preditiva de regressão com variáveis de entrada numéricas. As técnicas mais comuns são usar um coeficiente de correlação, como o de Pearson (Figura 13) para uma correlação linear ou métodos baseados em classificação para uma correlação não linear.

Figura 13 – Exemplo de Matriz de Correlação de Pearson, onde se avalia a correlação linear das variáveis entre si



Fonte: Fundação Cearense de Meteorologia e Recursos Hídricos (2015)

Ambas as correlações citadas consideram as variáveis como independentes, mas para o nosso trabalho de formatura buscamos uma métrica que compare o desempenho do acréscimo ou decréscimo de uma variável no modelo.

Buscando apenas utilizar variáveis que acrescentem informação para a regressão faremos uso do *F test* ou Teste F, que é um teste estatístico usado para comparar os modelos e verificar se a diferença entre eles é significativa.

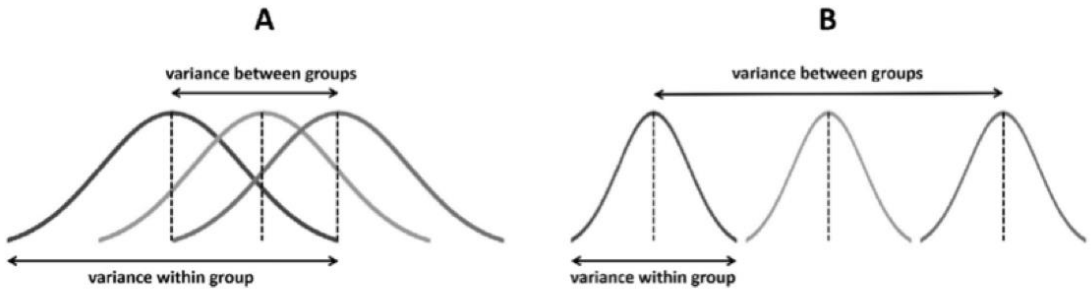
O Teste F faz um modelo de teste de hipótese X e Y, em que X é um modelo composto apenas por uma constante e Y é o modelo composto por uma constante e um recurso.

Os erros mínimos quadráticos em ambos os modelos são comparados e se verifica se a diferença de erros entre os modelos X e Y (ver Figura 14) é significativa ou introduzida por acaso. Para tanto calcularemos o valor de F que pode ser definido como:

Equação 3

$$F = \frac{\text{variação entre médias da amostra}}{\text{variação dentro das amostras}}$$

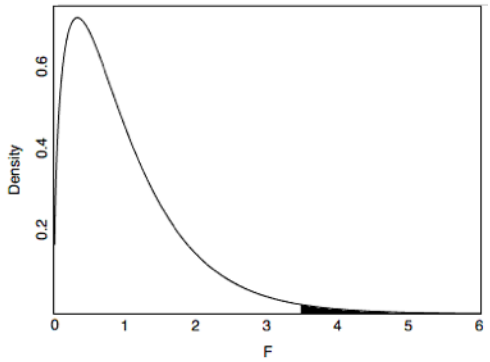
Figura 14 – Exemplos de Hipóteses do Teste F



Fonte: e-Handbook of Statistical Methods (1998)

A partir deste valor F compararemos com uma distribuição F, semelhante a ilustrada abaixo, na qual desempenharemos um teste de hipótese. A hipótese nula, área em branco da figura, representa que podemos assumir que a média dos modelos são iguais e, portanto, não há acréscimo de informação no modelo. A hipótese não nula, área negra do gráfico, não nos permite assumir a premissa da hipótese nula e, dessa forma, observamos um acréscimo de informação na regressão. Ver Figura 15

Figura 15 – Teste F



Fonte: e-Handbook of Statistical Methods (1998)

O teste F é útil na seleção de *features*, pois sabemos o significado de cada *feature* na melhoria do modelo.

Existem algumas desvantagens do uso do *F Test* para seleção de *features*. O teste F verifica e captura apenas relações lineares entre inputs e outputs. Um recurso altamente correlacionado recebe uma pontuação mais alta e os recursos menos correlacionados recebem uma pontuação mais baixa. A correlação pode ser altamente enganosa, pois não captura fortes relacionamentos não lineares.

3.6. Pré-Processamento: Escalonamento dos dados

Na maioria das vezes, o conjunto de dados contém variáveis de diferentes magnitudes, unidade e intervalo, porém, a maioria dos algoritmos de aprendizado de máquina utiliza a distância euclidiana como parâmetro, gerando problemas. Os resultados podem variar enormemente entre diferentes unidades, variáveis com maiores magnitudes poderiam ter um peso maior no cálculo. Assim, para evitar esse comportamento insatisfatório, utilizamos o escalonamento.

O *StandardScaler* (Equação 4) é um recurso que padroniza subtraindo a média e, em seguida, escalando para a variação da unidade. Variação unitária significa dividir todos os valores pelo desvio padrão. O *StandardScaler* não atende à definição estrita de escala que introduzimos anteriormente.

Equação 4

$$z = \frac{x_i - \mu}{\sigma}$$

O *StandardScaler* resulta em uma distribuição com um desvio padrão igual a 1. A variância também é igual a 1, porque a variância = desvio padrão ao quadrado.

3.7. Regressores

Segundo Izbicki e Santos (2019), as regressões representam um conjunto de técnicas que exploram a relação de uma variável independente com um grupo de outras variáveis, denominadas explicatórias, permitindo o ajuste de curvas e a estimação de valores com base em observações amostrais.

Entre os métodos mais usuais pode-se citar a Regressão Linear, que compreende um modelo no qual se busca definir uma equação linear que explique o comportamento de um conjunto de pontos observados, fazendo-se o uso do Método dos Mínimos Quadrados (IZBICKI e SANTOS, 2019).

Neste trabalho de formatura propõe-se um método que faz o uso de uma série de regressores, testando-os, a fim de se identificar aquele que se adequa da melhor maneira a um dado par Origem-Destino, fazendo uso das características socioeconômicas das regiões que compreendem a esse par, além da distância entre elas, a fim de estimar

possíveis valores faltantes da Matriz OD formada para a cidade de São Paulo com dados de radares rodoviários.

Os métodos utilizados são:

- Regressão Linear
- Regressão Ridge Bayesiana
- Support Vector Machine
- Algoritmo Passivo-Agressivo
- Least Angle Regression
- Stochastic Gradient Descent Regression

Uma descrição detalhada de cada um dos regressores utilizados neste trabalho de formatura pode ser encontrada no Anexo B

3.8. Otimização dos Hiperparâmetros

Os parâmetros do modelo são as variáveis que o método de aprendizado de máquina escolhe usar para ajustar os dados.

Por exemplo, para regressões lineares cada variável de entrada possui um peso associado que informa ao modelo o impacto que cada uma tem na predição final, de tal modo que a equação (Equação 5) que representa a regressão pode ser descrita como:

Equação 5

$$\hat{y}_i = \sum_{j=0}^{m} X_{ij} w_j$$

Onde observamos os valores de entrada **X** e seus pesos associados **w**. Esses pesos w são um exemplo dos parâmetros do modelo.

De muitas formas, esses parâmetros são o modelo. Ou seja, são eles que diferenciam seu modelo específico de outros modelos do mesmo tipo, que trabalham com dados semelhantes.

Os hiperparâmetros são variáveis que controlam o próprio processo de treinamento. Por exemplo, faz parte da configuração de uma rede neural profunda (não utilizada neste trabalho) decidir quantas camadas ocultas de nós precisam ser usadas entre a camada de entrada e a camada de saída, bem como quantos nós cada camada precisa usar. Essas variáveis não estão diretamente relacionadas aos dados de treinamento. Elas são variáveis de configuração. Os parâmetros mudam durante a execução de um treinamento, enquanto os hiperparâmetros geralmente permanecem constantes durante uma execução.

Os parâmetros do modelo são otimizados (ou seja, "ajustados") pelo processo de treinamento. Os hiperparâmetros são ajustados por meio da execução de todo o trabalho de treinamento, a observação da precisão agregada e o ajuste. Nos dois casos, você está modificando a composição do modelo tentando encontrar a melhor combinação para lidar com o problema.

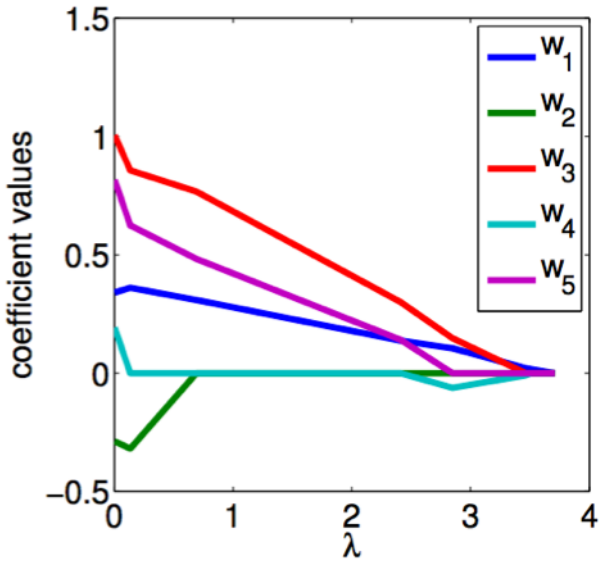
A regressão Lasso, utilizada neste trabalho, servirá de exemplo para ilustrarmos como a utilização de hiperparâmetros pode afetar os resultados da regressão. O erro associado a regressão lasso pode ser descrito pela seguinte equação:

Equação 6

$$\sum_{i=1}^{M} (y_i - \hat{y_i})^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} |w_j|$$

Assim como a regressão os parâmetros são os pesos W. Dentre os hiperparâmetros, podemos observar na equação acima o valor **λ**, que permanece o mesmo durante todo o processo de treinamento. Esse valor multiplica o peso residual, que representa uma penalidade ao modelo impondo-se uma margem de erro que não depende dos valores de entrada. Por exemplo, o gráfico abaixo nos da uma ideia de como o hiperparâmetro **λ** pode afetar os parâmetros W e, portanto, o resultado da regressão.

Figura 16 – Comportamento dos Coeficientes com a Variação do Hiperparâmetro



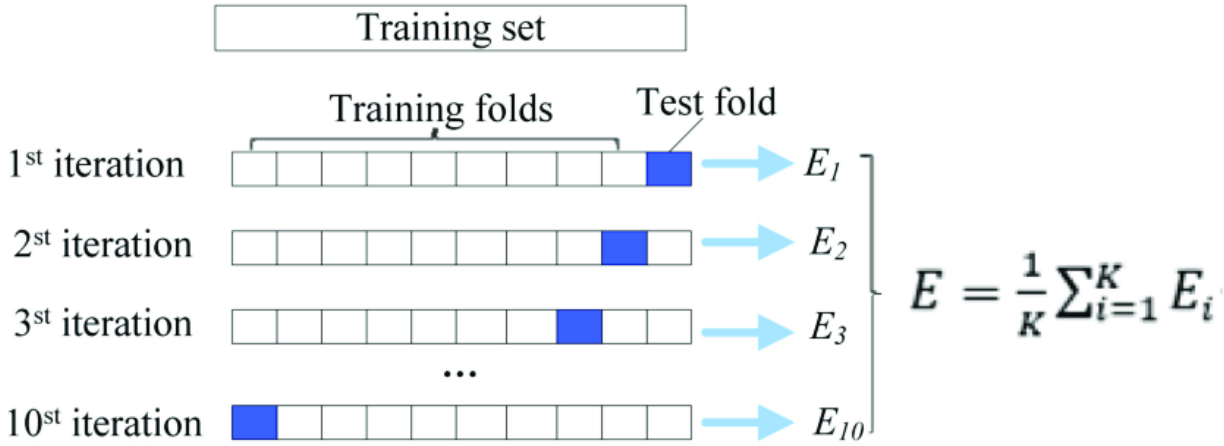
Fonte: Cross Validated (2019)

Para executarmos esta otimização em nossas regressões, de modo computacionalmente viável, utilizaremos o método *Grid Search* ou Pesquisa de Grade, que examina cada combinação de hiperparâmetros. Isso significa que todas as combinações de valores de hiperparâmetros especificados serão exploradas.

O ajuste de hiperparâmetros otimiza uma única variável, também chamada de métrica de hiperparâmetro, especificada durante o algoritmo. No nosso trabalho essa métrica é o R^2 score.

Para obter o melhor desempenho (performance) é feita uma comparação eficiente através do algoritimo *GridSearchCV*, que executa a comparação através de uma validação cruzada (*cross-validation*). E, utiliza ainda uma parte do conjunto de dados (*dataset*) de treinamento como *conjunto de dados* de validação. Dessa forma, o algoritmo executa um número arbitrário de iterações, onde há uma combinação diferente de *conjunto* de treinamento e validação. A métrica é obtida como a média entre essas combinações. Ver Figura 17

Figura 17 – Otimização de Hiperparâmetros



Fonte: ROASEN. K. (2016)

Segue (Tabela 6) os valores que trabalharemos para otimização de hiperparâmetros. Esses valores foram obtidos com base observações manuais e podem ser melhorados.

Tabela 6 – Hiperparâmetros por regressão

Regressão	Hiperparametro	Valores
Stochastic Gradient Descent Regression	alpha	1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3
	max iterations	1000, 500,500
	learning rate	constant,optimal,invscaling,adaptive
	penalty	nenhum, l2, elasticnet
	loss	squared loss,huber,epsilon insensitive,squared epsiloninsensitive
Lasso	alpha	0.005, 0.02, 0.025, 0.03, 0.05, 0.06
Ridge	alpha	200, 230, 250, 265, 270, 275, 290, 300, 550, 580, 600, 620, 650
ARDR	alpha	1e-7,1e-6, 1e-5, 1e-4, 1e-3, 1e-2
	lambda	1e-7,1e-6, 1e-5, 1e-4, 1e-3, 1e-2
Passive Agressive Regression		100 , 500 , 1000 , 20000

Fonte: Elaboração própria (2020)

3.9. Curvas de Aprendizagem

O desempenho de um modelo de aprendizado de máquina é considerado bom com base em sua previsão e em quão bem pode ser generalizado um conjunto de dados de teste independente. Com base no desempenho de diferentes modelos, escolhemos o modelo com melhor desempenho.

Existem dois principais cenários que busca-se evitar quando se trata de complexidade do modelo:

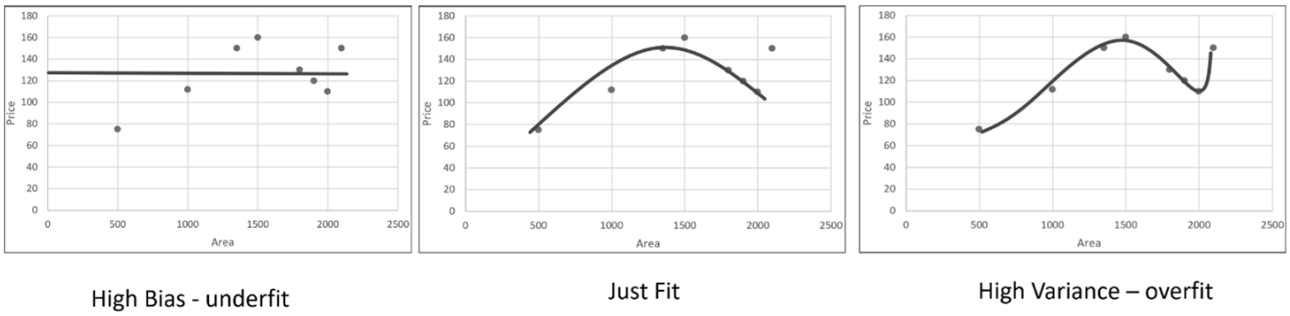
- ***Overfitting:***

Um modelo é dito como sobreajustado (*overfitted*) quando ele não generaliza bem, ou seja, ele só faz boas predições para aquele determinado conjunto de dados para qual ele foi treinado ou dados muito parecidos. Em geral esses modelos são mais complexos, polinômios de graus elevados, que conseguem capturar o ‘ruído’ dos dados.

- ***Underfitting***

Diz-se que um modelo estatístico ou um algoritmo de aprendizado de máquina ter underfitted quando não é possível capturar a tendência subjacente dos dados. Em geral são modelos poucos complexos. Ver Figura 18

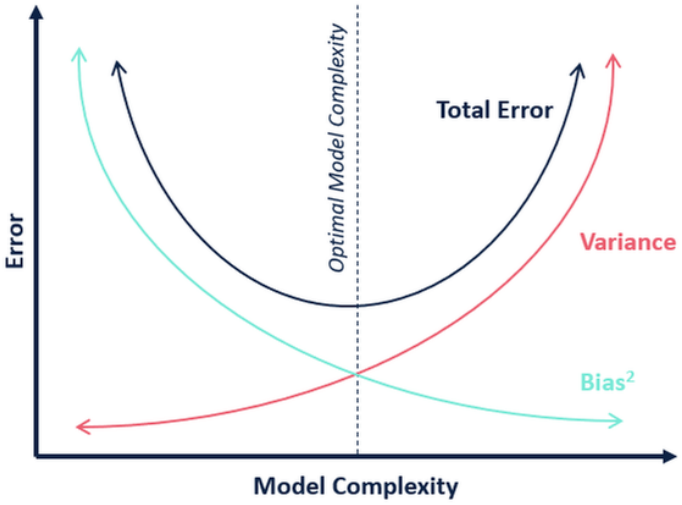
Figura 18 – Exemplos de diferentes comportamentos de modelos



Fonte: Towards Data Science (2018)

A partir disso, pode-se estabelecer que há um ponto ótimo de complexidade onde o modelo consegue generalizar bem sem cair nos dois cenários citados acima. Ver Figura 19

Figura 19 – *Trade-off* entre *bias* e variância



Fonte: Towards Data Science (2018)

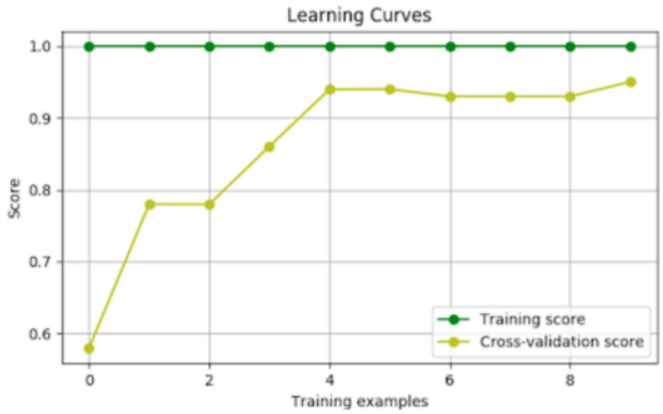
Faz-se importante então verificarmos que o regressor aplicado não ultrapasse este ponto ótimo de complexidade sobre o risco de não generalizar bem e, portanto, não servir para os propósitos deste trabalho de formatura.

Para tanto, usamos da mesma solução encontrada para otimização dos hiperparâmetros: *cross* validação. A *cross* validação será desempenhada agora não para obter uma métrica final, mas para acompanhar a tendência das métricas através das curvas de aprendizagem. Ver Figura 20

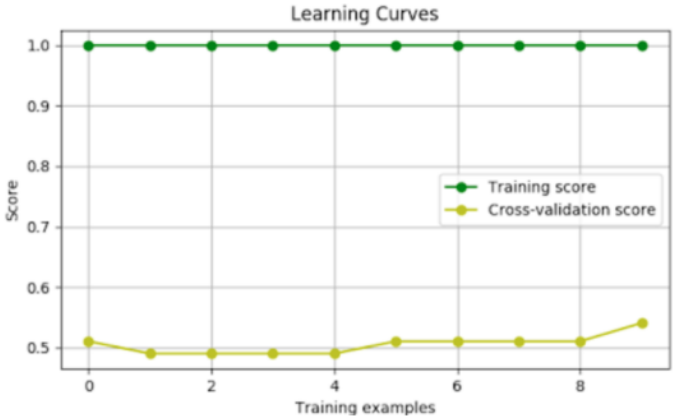
Figura 20 – I. Cenário com underfitting, II. Cenário ideal, III. Cenário com overfitting

I.

II.



III.



Fonte: UDACITY (2019)

A convergência das curvas, tanto do conjunto de validação quanto do conjunto de teste – conjuntos selecionados aleatoriamente - demonstra que há uma generalização dos dados e, portanto, torna-se requisito obrigatório para seleção dos possíveis regressores a serem aplicados a determinado conjunto de dados (*dataset*).

4. RESULTADOS OBTIDOS

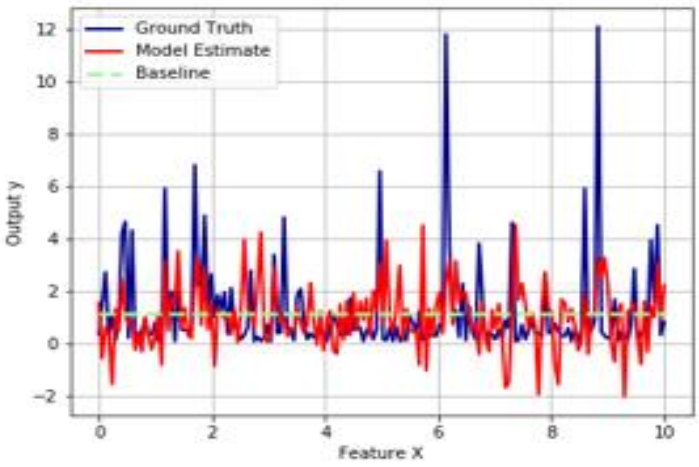
4.1. Exemplo: Resultados de Um Radar Específico

Para ilustramos nossos resultados, demonstraremos a partir de um exemplo o comportamento da amostra em três cenários:

- Resultados iniciais fornecidos ('*Ground Truth*').
- Resultados obtidos a partir do modelo de regressão.
- Média dos valores ('*baseline*').

Para propósitos de visualização buscamos fazer a disposição destes resultados transformando um espaço de 11 dimensões em um espaço bidimensional onde o eixo das abcissas representa os vetores de cada um dos pontos inicialmente dispostos em um espaço em 11 dimensões, dessa forma cada um dos vetores possui informações das variáveis de entrada de cada ponto. Desta maneira a comparação dos resultados que queremos ilustra pode ser demonstrada como:

Figura 21 – Resultados obtidos para um radar



Fonte: Elaboração Própria (2020)

Como podemos observar, em vermelho encontram-se os valores obtidos pelo modelo, e em verde-claro encontra-se a média dos valores fornecidos. À medida que há melhoria na métrica a linha vermelha passa a se aproximar da linha azul e, em um modelo perfeito a curva vermelha conseguiria ser igual a curva azul. Com este gráfico, conseguimos analisar de uma forma prática o quão este modelo consegue estimar melhor que a média. Modelos com métricas (R^2 score) negativas apresentam curvas que destoam dos dados fornecidos a ponto de gerarem um erro maior que a média.

Para o modelo ser considerado para posterior comparação entre métricas ele deve possuir curvas de aprendizagem convergentes, como discutido no capítulo anterior. Neste capítulo demonstraremos junto com os resultados, as curvas de regressão. Desta forma, conseguimos filtrar as soluções que devem ser consideradas para comparação entre modelos.

Utilizaremos como exemplo para visualização do processo para chegar aos resultados, o conjunto de dados correspondente aos pares Origem-Destino, cujo radar de origem é identificado pelo número 12716.

O radar apresenta 116 volumes faltantes havendo como principais varáveis correlacionada aos seus volumes as seguintes:

- Custo do trajeto,
- Renda Familiar Média,
- Renda per capita,
- Profissionais liberais,
- Emprego em serviços – outros,
- Produção de viagens por taxi convencional,
- Produção de viagens por moto,
- Transporte coletivo,
- Atração de viagens por moto

Cada um destes dados está descrito em detalhes no Anexo A.

Com estes dados de entrada, separamos a amostra em treinamento e teste. Na sequência, aplicamos cada um dos regressores, otimizando seus hiperparâmetros, conforme detalhado no item 3.8

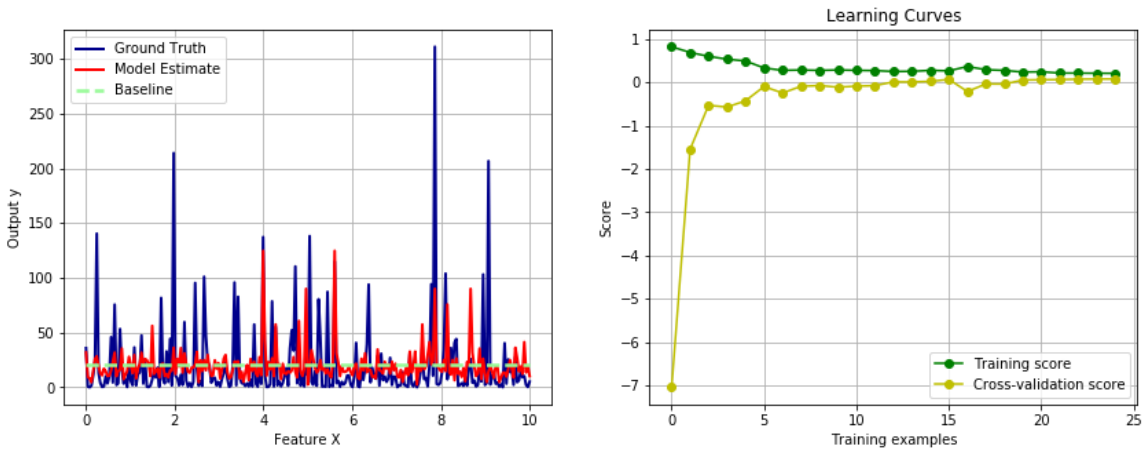
Cada um dos regressores, citados abaixo, possui uma explicação mais detalhada sobre seu funcionamento no Anexo B.

Obtemos, então, os seguintes resultados para o exemplo:

I. Regressão Passiva Agressiva:

Figura 22 – Regressor Passivo Agressivo

```
Pipeline(memory=None,
         steps=[('standard',
                 StandardScaler(copy=True, with_mean=True, with_std=True)),
                ('fit',
                 PassiveAggressiveRegressor(C=1.0, average=False,
                                            early_stopping=False, epsilon=0.1,
                                            fit_intercept=True,
                                            loss='epsilon_insensitive',
                                            max_iter=20000, n_iter_no_change=5,
                                            random_state=None, shuffle=True,
                                            tol=0.001, validation_fraction=0.1,
                                            verbose=0, warm_start=False))],
         verbose=False)
```



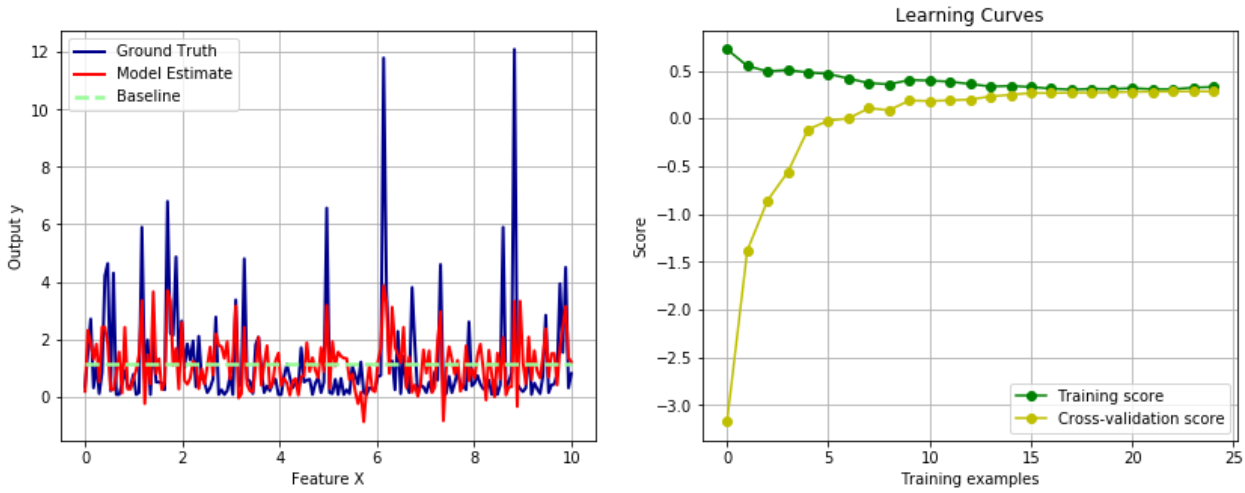
Fonte: Elaboração Própria (2020)

Métrica R^2: 0.24

II. Regressão Lasso:

Figura 23 – Regressor Lasso

```
Pipeline(memory=None,
         steps=[('standard',
                 StandardScaler(copy=True, with_mean=True, with_std=True)),
                ('fit',
                 Lasso(alpha=0.06, copy_X=True, fit_intercept=True,
                       max_iter=1000, normalize=False, positive=False,
                       precompute=False, random_state=None, selection='cyclic',
                       tol=0.0001, warm_start=False))],
         verbose=False)
```



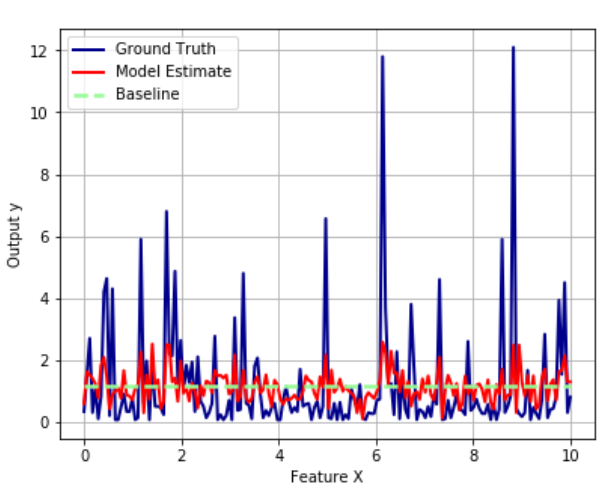
Fonte: Elaboração Própria (2020)

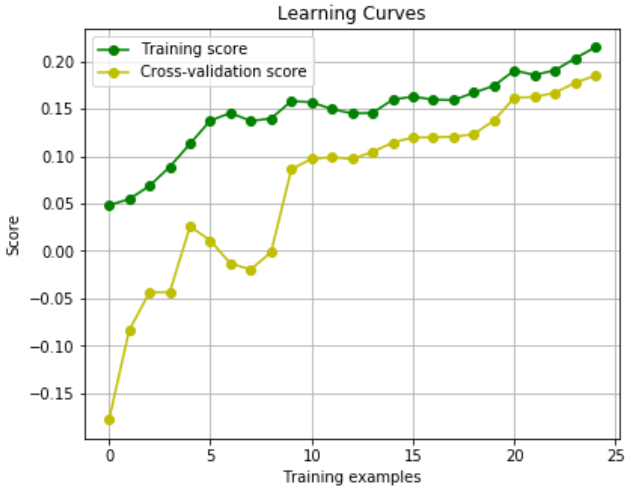
Métrica R^2: 0.36

III. Regressão Ridge:

Figura 24 – Regressão Ridge

```
Pipeline(memory=None,
         steps=[('standard',
                 StandardScaler(copy=True, with_mean=True, with_std=True)),
                ('fit',
                 Ridge(alpha=200, copy_X=True, fit_intercept=True,
                       max_iter=None, normalize=False, random_state=None,
                       solver='auto', tol=0.001))],
         verbose=False)
```





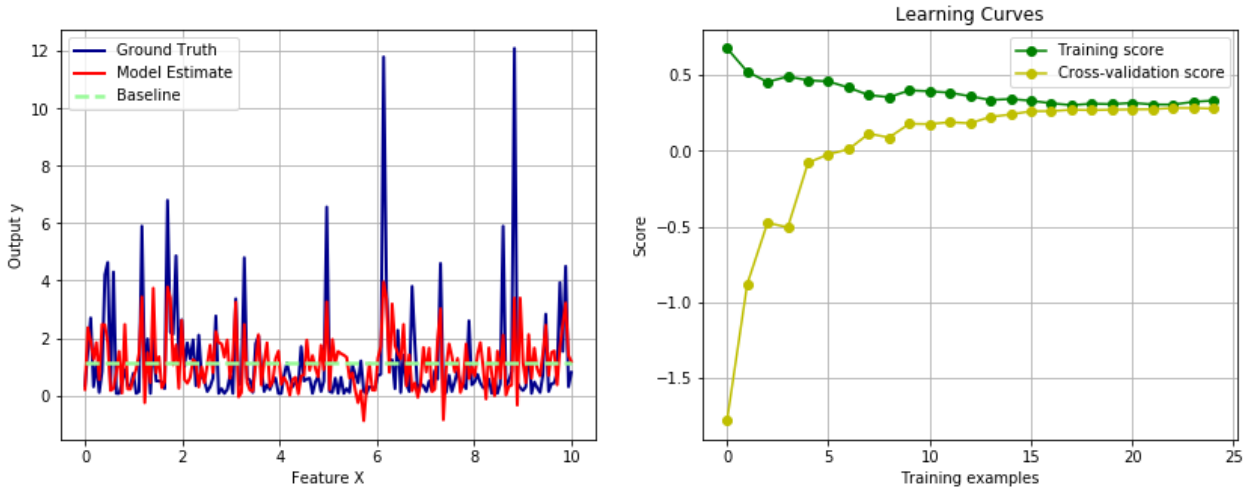
Fonte: Elaboração Própria (2020)

Métrica R^2: 0.25

IV. Regressão Gradiente Descendente Estocástico:

Figura 25 – Stochastic Gradient Descent Regression

```
Pipeline(memory=None,
         steps=[('standard',
                 StandardScaler(copy=True, with_mean=True, with_std=True)),
                ('fit',
                 SGDRegressor(alpha=0.1, average=False, early_stopping=False,
                              epsilon=0.1, eta0=0.01, fit_intercept=True,
                              l1_ratio=0.15, learning_rate='invscaling',
                              loss='squared_loss', max_iter=1000,
                              n_iter_no_change=5, penalty='l1', power_t=0.25,
                              random_state=None, shuffle=True, tol=0.001,
                              validation_fraction=0.1, verbose=0,
                              warm_start=False))],
         verbose=False)
```



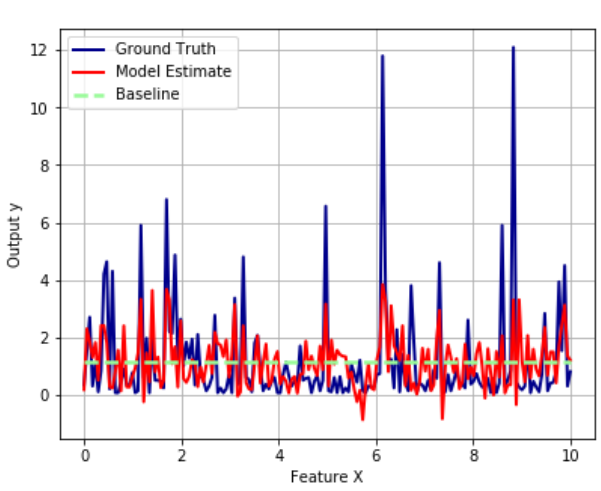
Fonte: Elaboração Própria (2020)

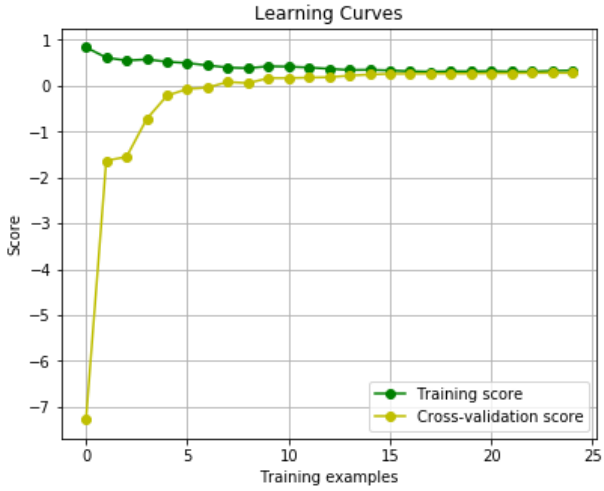
Métrica R^2: 0.36

V. Regressão Lasso LARS:

Figura 26 – Lasso LARS

```
Pipeline(memory=None,
         steps=[('standard',
                 StandardScaler(copy=True, with_mean=True, with_std=True)),
                ('fit',
                 LassoLars(alpha=0.005, copy_X=True, eps=2.220446049250313e-16,
                           fit_intercept=True, fit_path=True, max_iter=500,
                           normalize=True, positive=False, precompute='auto',
                           verbose=False))],
         verbose=False)
```





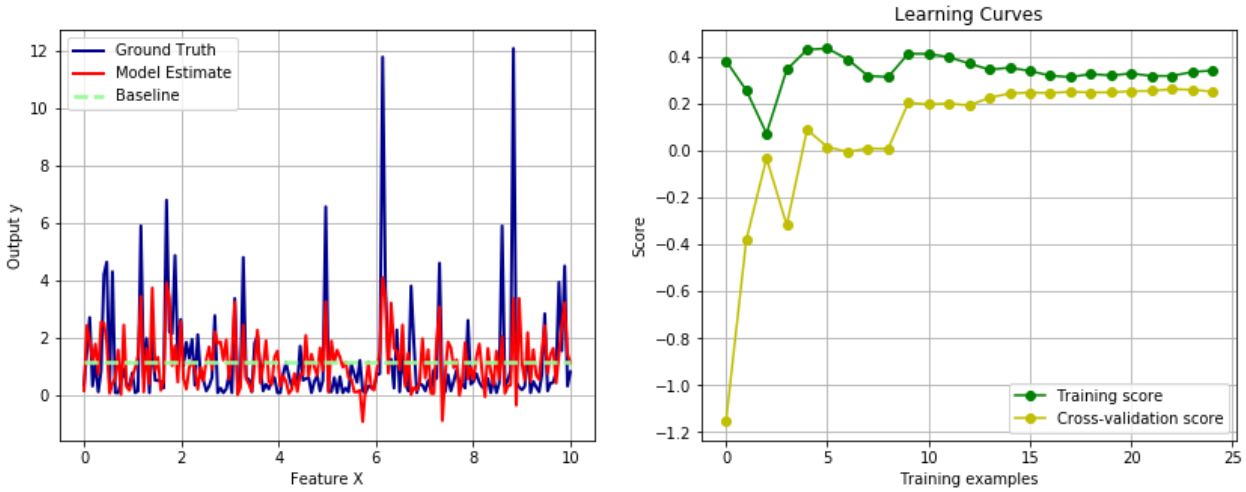
Fonte: Elaboração Própria (2020)

Métrica R^2: 0.35

VI. Regressão Ridge Bayesiana:

Figura 27 – Regressão Ridge Bayesiana

```
Pipeline(memory=None,
         steps=[('standard',
                 StandardScaler(copy=True, with_mean=True, with_std=True)),
                ('fit',
                 BayesianRidge(alpha_1=550, alpha_2=650, compute_score=False,
                               copy_X=True, fit_intercept=True, lambda_1=1e-06,
                               lambda_2=1e-06, n_iter=300, normalize=False,
                               tol=0.001, verbose=False))],
         verbose=False)
```



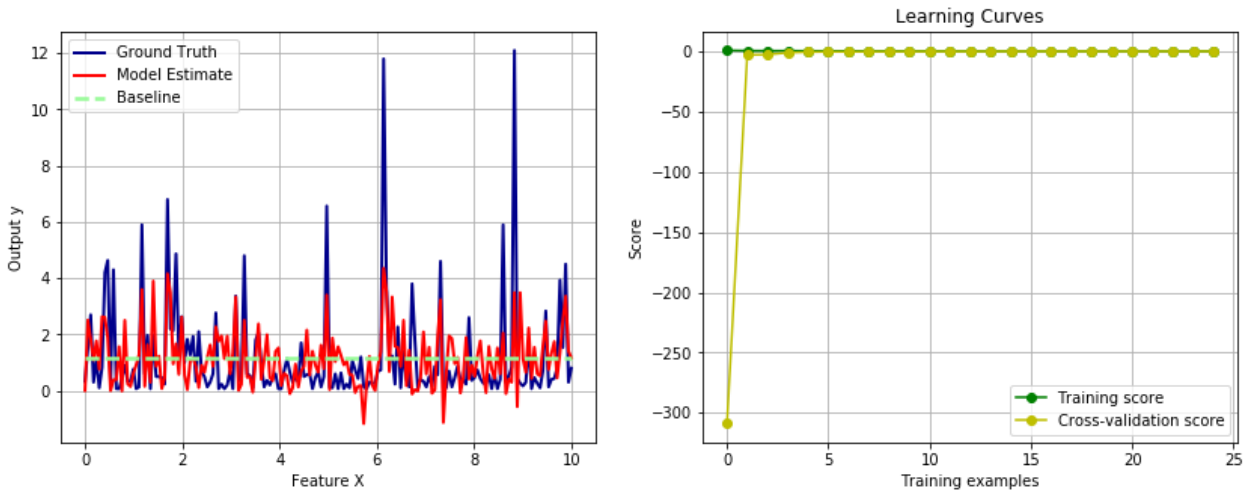
Fonte: Elaboração Própria (2020)

Métrica R^2: 0.37

VII. Regressão Linear:

Figura 28 – Regressão Linear

```
Pipeline(memory=None,
         steps=[('fit',
                 LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
                                  normalize=True))],
         verbose=False)
```



Fonte: Elaboração Própria (2020)

Métrica R²: 0.38

Tabela 7 – Resumo dos resultados obtidos para um radar (exemplo)

Radar 12.716

Regressão	Convergência	R^2
Linear	✓	0,38
Ridge Bayesiana	✓	0,37
Lasso	✓	0,36
Stochastic Gradient Descent	✓	0,36
Least Angle	✓	0,35
Ridge		0,25
Passiva Agressiva		0,24

Fonte: Elaboração Própria (2020)

Como podemos observar, alguns dos regressores aplicados serão descartados da comparação por não haver convergência das curvas de aprendizagem. São eles:

- Regressor Passivo agressivo
- Regressor Ridge

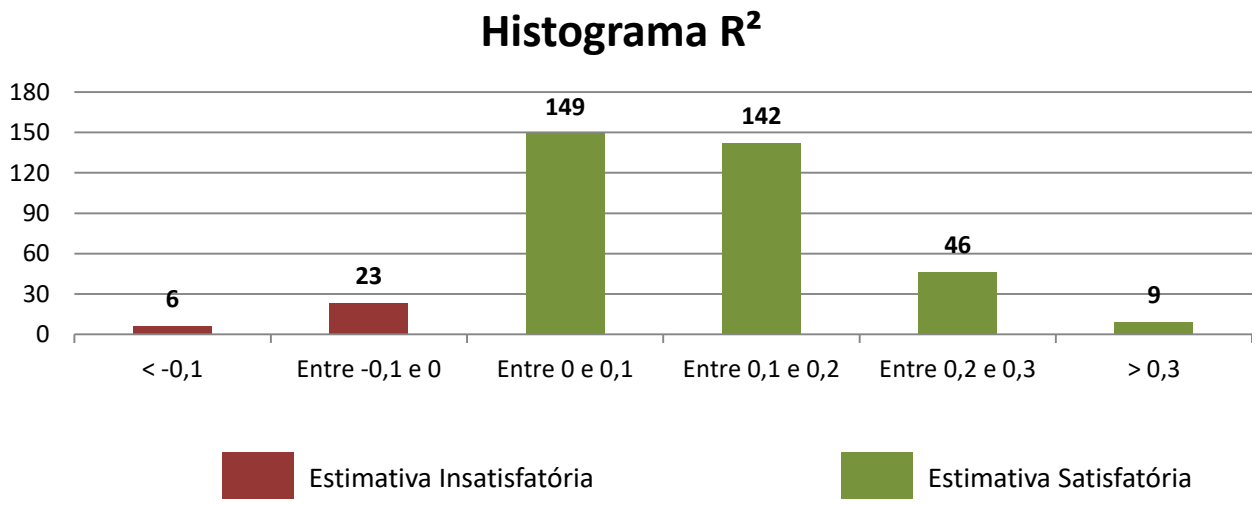
Havendo convergência escolhemos o regressor com melhor desempenho segundo a métrica R^2 score: nesse caso o Regressor Linear. Ver 7

Vale ressaltar que nenhum dos regressores apresentou métricas acima de 40%. Indicativo que, apesar do modelo estimar melhor que a média, ainda não o faz de maneira tão satisfatória.

4.2. Resultados Gerais do Trabalho

Agora, analisando não apenas um radar, mas sim os resultados de todo o conjunto de radares, embora não se tenha obtido altas correlações - como esperado, pode-se afirmar que os resultados obtidos por esse trabalho de formatura são promissores, uma vez que, apenas 7,7% das estimativas de valores faltantes podem ser avaliadas como piores do que estimar diretamente pela média, conforme o histograma da Figura 29

Figura 29 – Distribuição dos R^2 obtidos



Fonte: Elaboração Própria (2020)

Os resultados demonstram a capacidade que os modelos têm para gerar estimativas que se aproximem da realidade, porém neste trabalho de formatura as métricas demonstraram ser insuficientes, gerando questionamento sobre a capacidade dos dados obtidos em prover informação suficiente para alimentar os modelos, de modo a gerar resultados próximos da realidade.

Destaca-se, ainda, que os resultados completos obtidos por este trabalho de formatura (incluindo a matriz OD completa e todas as métricas obtidas por radar) podem ser acessados no Google Drive criado pelo grupo, cujo link se encontra no item 7 deste documento.

5. CONCLUSÕES

O uso de sistemas inteligentes de transportes é uma realidade que tem um grande potencial para otimizar a tomada de decisão na engenharia de tráfego e no planejamento urbano de maneira geral. Os sistemas de identificação automática de veículos são uma tecnologia que já traz grandes avanços não só para a fiscalização do cumprimento das normas de trânsito, mas também no monitoramento das condições da via, com informações de fluxo e volume, por exemplo.

Muito se pode evoluir com o avanço tecnológico e com o crescente emprego de ciência de dados na tomada de decisão estratégica em nível urbano e intermunicipal, visto que há uma imensa quantidade de dados a serem explorados que não foram analisados até agora e que podem ser interpretados de diversas maneiras, expondo o funcionamento e comportamento da cidade e de seus habitantes.

Este trabalho de formatura buscava, a partir da informação que está disponível hoje, fazer uso dessa tecnologia e estimar uma matriz OD, que tivesse seus *missing values* preenchidos por meio de um algoritmo que estimasse o melhor valor, com base em dados como número de empregos, renda média, população entre outros. Pode-se afirmar que esse objetivo foi alcançado, uma vez que, a maioria dos valores faltantes pôde ser estimada de forma satisfatória, conforme a Figura 29.

Todavia, a análise dos resultados alcançados com o desenvolvimento deste Trabalho de Formatura permite afirmar que dados socioeconômicos tem representatividade menor que a esperada no que tange a caracterização dos padrões de fluxos entre regiões de uma cidade, sendo pouco eficientes para a elaboração da matriz OD completa que se desejava. Por outro lado, as informações relativas ao custo estimado da viagem impactam as estimativas de maneira bem mais relevante, sendo a principal variável de entrada do modelo.

Outro ponto levantado pela equipe é o não uso das características da malha viária disponível, como o número de faixas na via e a sua capacidade, os quais também influenciam a dinâmica do tráfego rodoviário.

Portanto, é esperado que o uso de dados com maior granularidade e com características mais representativas - na geração de fluxos de viagens – possivelmente poderão trazer resultados mais satisfatórios para o uso de tecnologias já empregadas na infraestrutura de trânsito e, assim, se estimar, de maneira confiável, de onde as pessoas vêm e para onde elas vão dentro da cidade.

6. CONSIDERAÇÕES FINAIS

Mesmo obtendo-se um resultado satisfatório, é sugerido que estudos futuros incorporem às análises outros dados, como características da via e sua capacidade, dados de contagem e de pedágios, entre outros pertinentes ao fluxo de veículos. Outra sugestão é se utilizar de regressores e técnicas mais complexas, como árvore de decisão e rede neural, além de explorar e realizar uma otimização mais profunda de todos os hiperparâmetros relacionados a cada uma das regressões. Espera-se que com mais dados disponíveis e técnicas mais complexas de aprendizado de máquina, o modelo se comporte melhor e obtenha resultados mais próximos da realidade.

Além disso, este trabalho de formatura deixa como legado um modelo de aprendizado de máquina (“*machine learning*”) “simples” que pode ser utilizado amplamente. Tanto os dados relacionados aos fluxos e volumes (que neste trabalho se desejava estimar), quanto os dados socioeconômicos (que neste trabalho serviram como base das estimativas) são meras entradas do modelo e podem ser facilmente substituídas por outros tipos de dados. O modelo gerado por este trabalho de formatura interpreta os dados faltantes a serem imputados, a partir de dados conhecidos e de sua relação com os dados faltantes, quaisquer que sejam esses dados. Acredita-se, inclusive, que o modelo possa ser utilizado para estimar “*missing values*” de bases de dados de outras áreas do conhecimento, como: medicina, computação, administração, logística, entre outras.

Por fim, destaca-se que este trabalho de formatura não tem a pretensão de esgotar o assunto, mas sim evidenciar uma ampla área de aplicação para o aprendizado de máquina, sendo que diversas das análises propostas aqui podem continuar sendo estudadas e melhoradas significativamente.

7. LINKS PARA MATERIAIS PRODUZIDOS

Os resultados completos obtidos, bem como os códigos de programação do algorítimo construído encontram-se disponíveis nos seguintes links:

- https://github.com/alexcbfa/TCC2.git
- https://drive.google.com/open?id=1aVafrpZjzlcK-geK_Amx7u8pJpKW-8os

8. REFERÊNCIAS

BERNARDI, E. **Os sistemas de identificação veicular, em especial o reconhecimento automático de placas**. São Paulo: Universidade de São Paulo, 2015

CARDOSO, C. E. P. **Modelos tradicionais de transporte e tráfego**. São Paulo: Universidade de São Paulo, 2011.

CRAMMER, K., DEKEL, O., KESHET, J, SHALEV-SHWARTZ, S. e SINGER, Y. **Online Passive-Aggressive Algorithms**. Jerusalem, Israel: The Hebrew University, 2006

DANG, X., PENG, H., WANG, X. e ZHANG, H. **Theil-Sen Estimators in a multiple Linear Regression Model**. Estados Unidos: University of Mississippi e Yale University, 2009

EFRON, B., HASTIE, T, JOHNSTONE, I. e TIBSHIRANI, R. **Least Angle Regression**. Palo Alto: Stanford University, 2004

IZBICKI, R. e SANTOS, M. **Machine Learning sob a ótica da estatística – Uma abordagem preditivista para estatística com exemplos em R**. São Paulo: Universidade Federal de São Carlos e Insper Instituto de Ensino e Pesquisa, 2019.

LIITIAINEN, E. **Authomatic Relevance Determination**. Helsinki, Finlândia: Helsinki Univesity of Technology, 2006.

MARTINS, DOUGLAS F W CAPELOSSI. **Scalable method for origin-destination demand estimation using automatic vehicle identification data**. São Paulo: Universidade de São Paulo, 2019.

ORTÚZAR, J. de D.; WILLUMSEN, L. G. **Modelling Transport 2ª Edição**. Chischester, England, 1994.

PANAGIOTAKOPOULOS, C. TSAMPOUKA, P. **The Stochastic Gradient Descent for The Primal L1-SVM Optimization Revisited**. Grécia: Aristotle University of Thessaloniki, 2014.

PASANEN, L., HOLMSTROM, L e SILLANPAA, M. **Supporting Information for Bayesian LASSO, scale space and decision making in association genetics**. Oulu, Finlândia: University of Oulu, 2015.

WILHELM, F. **Explaining the Idea Behind Authomatic Relevance Determination and Bayesian Interpolation**. Amsterdam, 2016.

YALE UNIVERSITY DEPARTMENT OF STATISTICS. **Linear Regression,** New Haven, 1997. Disponível em: <http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>. Visitado em 20 de novembro de 2019.

Anexo A: Descrição das Variáveis Socioeconômicos Utilizadas

Tipo	Dado	Descrição
Dados Gerais	Domicilios	Representa o número de domicílios existentes em cada zona.
	Famílias	Representa o número de famílias que habitam cada zona.
	População	Representa o número de pessoas que habitam cada zona.
	Matrículas Escolares	Representa o número de alunos matriculados nas escolas de cada zona. Considerando Ensino Fundamental I corresponde ao período da 1ª à 4ª série do 1º grau (antigo Ensino Primário); Ensino Fundamental II corresponde ao período da 5ª à 8ª série do 1º grau (antigo Ginásio); Ensino Médio corresponde ao período da 1ª à 3ª (ou 4ª) série do 2º grau (antigo Colegial)
	Empregos	Representa o número total de postos de trabalho em cada zona.
	Automóveis	Representa o número total de automóveis detidos pelas pessoas que habitam cada zona.
	Viagens Produzidas	Representa o total de viagens produzidas por cada zona.
	Viagens Atraídas	Representa o total de viagens atraídas por cada zona.
	Área	Representa a área de cada zona em hectares.
Renda	Renda Total	Representa a soma da renda de totas as pessoas que habitam cada zona.
	Renda Média Familiar	Representa a média de renda familiar de cada zona. ***Renda Média Familiar = Renda Total / Famílias***
	Renda per Capita	Representa a média de renda por habitante de cada zona. ***Renda per Capita = Renda Total / População***
Automóveis	Automóveis por Família	Representa o número médio de automóveis por família de cada zona. ***Automóveis por Família = Automóveis / Famílias***

Tipo	Dado	Descrição
Vínculo Empregatício	Assalariado com Carteira	Pessoa que tem vínculo empregatício com empresa do setor público ou privado, possui jornada definida e que está registrada em carteira de trabalho em conformidade com a CLT.
	Assalariado sem Carteira	Pessoa que trabalha em empresa do setor público ou privado, possui jornada definida e que não tem contrato de trabalho formalizado pela CLT, ou seja, sem registro em carteira.
	Funcionário Público	Pessoa civil ou militar que trabalha em instituição pública (governo municipal, estadual, federal, empresa de economia mista, autarquia, fundações, etc.), contratada pelo regime do funcionalismo público (CLF), ou seja, na condição de funcionário público estatutário
	Autônomo	Pessoa que trabalha por conta própria, podendo desenvolver sua atividade ou prestar seus serviços para uma ou mais empresas ou para a população em geral.
	Empregador	Pessoa que é proprietária de uma empresa e que tem pelo menos um empregado contratado de forma permanente.
	Profissional Liberal	Pessoa de nível superior que exerce uma atividade ligada à sua formação universitária e que trabalha por conta própria para várias empresas ou para a população em geral, podendo ter até dois empregados.
	Dono de Negócio Familiar	Pessoa que gerencia um negócio ou uma empresa de sua propriedade exclusiva ou em sociedade com parentes. Normalmente nesse tipo de negócio só trabalha mão-de-obra familiar não remunerada, podendo ter até dois empregados remunerados.
	Trabalhador Familiar	Pessoa que trabalha em negócios de família sem remuneração salarial.

Tipo	Dado	Descrição
Condição de Ocupação	Trabalho Regular	Quando a pessoa declara ter um ou mais trabalhos que lhe garanta uma remuneração em dinheiro, independente da formalização do vínculo de trabalho. Inclui-se nesta situação também a pessoa que, no momento da pesquisa, não esteja trabalhando porque se encontra em gozo de férias ou por qualquer outro motivo que provoque interrupção temporária de suas atividades (greve, falta de matéria-prima etc.). Não devem ser considerados os afastamentos por licença médica que serão identificados, nesta questão, com a terceira opção. Deve ser também incluído como "Tem trabalho regular" o indivíduo que trabalha em negócio de parentes sem remuneração salarial.
	Faz Bico	Quando a pessoa declara que a atividade que está desenvolvendo é esporádica e os ganhos são avulsos.
	Em Licença Médica	Quando a pessoa tem um trabalho, mas está temporariamente afastada porque se encontra em licença médica para tratamento de saúde, independentemente do tempo de afastamento de seu emprego ou trabalho.
	Aposentado / Pensionista	Refere-se à pessoa que não tem trabalho, mas possui ganhos de aposentadoria ou recebe pensão da previdência social.
	Sem Trabalho	Quando a pessoa declara que está desempregada, independentemente de estar ou não procurando trabalho.
	Nunca Trabalhou	Essa categoria abrange a pessoa que declara estar sem trabalho e que nunca trabalhou.
	Dona de Casa	Quando a pessoa declara que tem como atividade cuidar de afazeres domésticos, ou seja, cuidar da casa, filhos, marido. Muitas dessas pessoas se identificam como "do lar".
	Estudante	Quando a pessoa tem 5 anos ou mais e só estuda e não exerce nenhum tipo de trabalho.
Emprego por Setor de Atividade	Secundário	O Setor Secundário inclui os setores da economia que transformam produtos, como indústrias e construção. Este setor geralmente pega os produtos provindos do Setor Primário e os transformam, ao ponto de servirem para serem usados para outros negócios, exportados ou para serem consumidos por consumidores domésticos.
	Terciário	Setor terciário (também conhecido como setor de serviços) é aquele que engloba as atividades de serviços e comércio de produtos.
	Outros	Inclui outros setores da Economia.

Tipo	Dado	Descrição
Emprego por Classe de Atividade	Agrícola	Atividades agrícolas, de reflorestamento, pecuária e outras que envolvem criação de animais, além das atividades extrativas vegetal e de pesca.
	Construção Civil	Inclui as atividades de construção e reparação de edificações e obras de infraestrutura.
	Indústria	Atividade cujo produto passa por processo de transformação ou beneficiamento. Inclui as atividades relativas à extração mineral.
	Comércio	Atividade de vendas de mercadorias realizada diretamente ao consumidor (vendas a varejo) ou para as empresas (vendas por atacado). Pode realizar-se tanto em estabelecimentos fixos ou ambulantes (nas vias públicas) ou diretamente em visita ao cliente.
	Serviço de transporte de carga	Atividade de transporte de carga de mercadorias ou valores.
	Serviço de transporte de passageiros	Atividade de transporte de passageiros rodoviários, ferroviários, metroviários, aéreos e outros.
	Serviço creditício-financeiro	Atividade ligada aos serviços de crédito e financeiros (bancos, bolsa de valores), inclusive seguros e cartões de crédito.
	Serviço pessoal	Atividade de embelezamento pessoal; higiene; academia de dança, ginástica e luta; sauna, massagem e outros definidos como de necessidade pessoal.
	Serviço de alimentação	Atividade ligada ao fornecimento de alimentação em bares, padarias, restaurantes, lanchonetes, barracas e outros vendedores de rua; serviços de entrega a domicílio de alimentos para consumo imediato.
	Serviço de saúde	Atividade ligada aos serviços de saúde (hospitais, maternidades, consultórios, análises clínico-laboratoriais).
	Serviço de educação	Atividade ligada a todos os tipos de escola pública ou privada, e as atividades dos professores particulares.
	Serviço especializado	Atividade ligada aos serviços de escritório, de assessorias e consultorias técnicas, jurídicas, econômicas, contábeis, serviços de pesquisa, serviços de processamento, análise e programação de dados e outros serviços técnicos profissionais.
	Serviço de administração pública	Atividade vinculada aos Poderes Legislativo, Judiciário e Executivo; serviços administrativos federais, estaduais, municipais e autárquicos; Exército, Marinha e Aeronáutica; Polícia Militar e Civil; Corpo de Bombeiros; e outras organizações governamentais.
	Outros Serviços	Todos os demais setores não classificados anteriormente

Anexo B: Descrição dos Regressores Utilizados

B.1. Regressão Linear

A Regressão Linear representa um modelo no qual se busca definir uma equação linear que explique o comportamento de um conjunto de pontos observados. Para tal, deve-se determinar a correlação entre as variáveis que se deseja empregar no modelo, a fim de se concluir se elas são, de fato associáveis.

Para se definir, o equacionamento da função que explicará o modelo, é usual se empregar o método dos mínimos quadrados, o qual mede o distanciamento no eixo vertical de cada ponto ao seu correspondente na função definida.

O MMQ, para uma função linear e um dado conjunto de pontos $(\mathrm{x_i}, \mathrm{y_i})$, é definido como (Equação 7):

Equação 7

$$min \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Onde a e b são os coeficientes da reta $y = ax + b$ que define a regressão.

Para modelos em que existem mais de uma variável explicativa, tem-se a regressão linear de múltiplas variáveis, em que busca-se um equacionamento do tipo (Equação 8):

Equação 8

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k \beta x_{ki} + \varepsilon_i$$

Para se determinar os parâmetros estimadores de mínimos quadrados β_k, deve-se minimizar o erro quadrático geral (Equação 9).

Equação 9

$$\sum_{i=1}^{n}(\varepsilon_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \cdots - \beta_k x_{ki})^2$$

O qual é solucionado por meio da resolução de um sistema de equações parciais.

B.2. Regressão Ridge Bayesiana

A Regressão Ridge Bayesiana, faz uso do método estatístico de inferência de Bayes, no qual o teorema de Bayes é usado para atualizar as probabilidades para uma dada hipótese a medida que novas informações e dados se tornam disponíveis. Segundo Pasanen, Holmström e Silanpää (2015), tal regressão, é obtida ao se utilizar de uma regressão do tipo $\hat{\beta} = arg_{\beta}min\,(\left|\left|y - \mu 1 - \mathrm{X}\beta\right|\right|^2 + \lambda||\beta||^2$, em que quanto maior o valor de λ mais os componentes de β se aproximam de zero. A partir disso, se dá um caráter bayesiano ao modelo, considerando-se que o minimizador do problema denotado é a média posterior de um novo modelo denotado por $\beta_j = \mathrm{N}(0, \frac{\sigma^2}{\lambda})$, para todo *j*.

B.3. Support Vector Machine

No campo do conhecimento de aprendizado de máquina, uma Support Vector Machine (SVM) é um modelo de aprendizado supervisionado o qual faz uso de uma série de algoritmos para analisar dados e, assim, criar um modelo de classificação ou de regressão. Segundo Cristianini e Shawe-Taylor (2000), as SVMs constroem um hiperplano ou n hiperplanos em espaço vetorial de n dimensões. Geralmente, devido a

alta dificuldade de se traçar um limite linear entre os conjuntos de dados, o espaço original no qual os pontos observados estão localizados tem sua dimensão expandida, dando maior liberdade aos processos de segregação de pontos de diferentes conjuntos, os quais são definidos por meio de uma função de Kernel.

Com isso, os hiperplanos na dimensão maior são definidos como o conjunto de pontos cujo produto escalar com um vetor nesse plano é constante. O conjunto de vetores forma um conjunto ortogonal. Tais vetores podem ser parametrizados formando combinações lineares do conjunto imagem dos pontos presentes na base de dados original, permitindo que a soma das funções de Kernel represente a proximidade entre os pontos de teste dos pontos originais. Assim, a SVM permite com que a distância entre os pontos de uma amostra possa ser medida por meio de uma função relacionada a eles em um espaço auxiliar de dimensão maior que o original, onde não seria possível traçar um método de classificação.

B.4. Algoritmo Passivo-Agressivo

Para se definir um algoritmo, deve-se determinar como será iniciado o vetor de peso w w_1 e se definir a regra de atualização usada para modificá-lo no fim de cada passo. No modelo de algoritmo Passivo-Agressivo tal vetor é inicializado em (0, ..., 0) para todas as variáveis, porém cada uma delas possui uma regra de atualização diferente. Primeiro, deve-se focar na variável de regra mais simples, que no passo t define o novo vetor de peso w_{t+1} como a solução para o problema de otimização da Equação 10.

Equação 10

$$w_{t+1} = \frac{1}{2} \operatorname{argmin} \frac{1}{2} \| w - w_t \|^2, \ell(w;(x_t,y_t)) = 0, w \epsilon \mathbb{R}^n$$

w_{t+1} é definido como a projeção de w_t no semi-espaço vetorial que anula as perdas no exemplo dado. O algoritmo resultante é passivo sempre que a perda é zero, ou seja $w_{t+1} = w_t$. Em contraste, nas rodadas em que a perda é positiva, o algoritmo força agressivamente w_{t+1} a satisfazer a restrição $\ell(w_{t+1}; (x_t, y_t)) = 0$, independentemente do tamanho da etapa necessária.

De certa maneira, a atualização exige que w_{t+1} classifique corretamente o exemplo atual com uma margem suficientemente alta e, portanto, força-o a progredir. Todavia, o novo peso deve permanecer o mais próximo possível do anterior, mantendo, assim, as informações aprendidas previamente. Assim, a solução do método pode ser sintetizada na forma da Equação 11:

Equação 11

$$w_{t+1} = w_{t+}\tau_t y_t x_t, \text{ onde } \tau_t = {\ell_t}/{||x_t||^2}$$

O algoritmo Passivo-Agressivo faz uso de uma estratégia de atualização agressiva, modificando o tanto que for necessário o vetor de peso a fim de satisfazer as condições impostas, o que pode, em certas situações, distorcer os resultados.

B.5. Least Angle Regression

O algoritmo de Least Angle Regression (LARS) permite o uso de uma formulação matemática simples, porém bastante célere, sendo necessárias n etapas para se obter o conjunto completo de soluções, em que n é o número de covariáveis do modelo.

Para se definir esse tipo de regressão, deve-se começar com todos os coeficientes iguais a zero, a fim de se encontrar o preditor mais correlacionado com a resposta. Dá-se o maior passo possível na direção desse preditor até o ponto em que outro

preditortenha tanta correlação quanto o atual. Nesse ponto, passa-se a seguir em uma direção equiangular entre os dois preditores até que uma terceira variável chegue ao nível de correlação das duas primeiras,e prossegue de forma equiangular entre as três, até que uma quarta variável entre, e assim por diante. Assim, a LARS compõem estimadores μ = Xβ, sucessivamente, adicionando, a cada passo, uma nova covariável.

B.6. Stochastic Gradient Descent Regression

Na Stochastic Gradient Descent Regression há uma redução gradiente em relação à função objetivo na qual o risco empírico $\mathbf{1}/\boldsymbol{m}\sum_{\boldsymbol{k=1}}^{\boldsymbol{m}}\mathbf{max}\{\mathbf{0},\mathbf{1}-\boldsymbol{w}\times\boldsymbol{y_k}\}$ é aproximado pelo risco instantâneo $\mathbf{max}\{\mathbf{0},\mathbf{1}-\boldsymbol{w}\times\boldsymbol{y_k}\}$ de uma única observação. A regra geral de atualização desse tipo de regressor é definida pela Equação 12

Equação 12

$$\boldsymbol{w_{t+1}} = \boldsymbol{w_t} - \boldsymbol{\eta_t}\boldsymbol{\nabla w_t}\left[\frac{\mathbf{1}}{\mathbf{2}}||w_t||^2 + \frac{1}{\lambda}\max\{0,1 - w_t \cdot y_k\right]$$

Onde η_té a taxa de aprendizado e ∇w_t representa um subgradiente em relação ao peso, uma vez que a perda da margem flexível de norma 1 é apenas diferenciável por partes (t ≥ 0). Deve-se, então, escolher uma taxa de aprendizado η_t = 1 / (t + 1) que satisfaça as condições $\sum_{t=0}^{\infty}{\eta_t}^2 < \infty$ e $\sum_{t=0}^{\infty}\eta_t = \infty$, geralmente, impostas na análise de convergência de aproximações estocásticas.

Então, percebendo-se que $w_t - \frac{1}{t+1}w_t = \frac{1}{t+1}w_t$, obtém-se a atualização (Equação 13)

Equação 13

$$w_{t+1} = \frac{t}{t+1} w_t + \frac{1}{\lambda(t+1)} y_k$$

toda vez que $w_t \cdot y_k \leq 1$ e $w_{t+1} = \frac{t}{t+1} w_t$

Ao se derivar a regra de atualização acima, escolhe-se $wt - \lambda^{-1} y_k$ para o subgradiente no ponto $w_t \cdot y_k = 1$ onde a perda de margem flexível de norma 1 não é diferenciável.

Assim, assume-se que $w_0 = 0$, e percebe-se que, se$w_t \cdot y_k > 1$, a atualização consiste em um encolhimento puro do vetor de peso atual pelo fator t / (t + 1).