UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

FÁBIO INNOCENTE ALVES

Predição de demanda para sistemas de bicicletas compartilhadas com estações utilizando agregação de dados meteorológicos

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Prof. Dr. Luciana Buriol

Porto Alegre
2019

“Amor Fati: amor ao destino”

— EXPRESSÃO LATINA

AGRADECIMENTOS

Dedico meus sinceros agradecimento,

Aos meus pais, a minha família, a todo pagador de impostos e a todos servidores desta instituição por me darem a oportunidade de cursar uma universidade de excelência.

Agradeço à Luciana Buriol, minha orientadora, pelo apoio e instrução para a realização deste trabalho e meu desenvolvimento no ecossistema de inovação.

Finalmente, gostaria de agradecer a mim por não ter desistido e fazer bom uso do conhecimento que me foi compartilhado todos esses anos.

RESUMO

Startups são empresas jovens de base tecnológica. Para serem enxutas e competitivas, é importante se manterem essencialistas e eficientes em suas atividades. Em sistemas de compartilhamento de bicicletas com estações podemos verificar práticas eficientes em operações de balanceamento que maximizam a utilização pelos usuários. Para isso, é necessário que o sistema ofereça alta disponibilidade de bicicletas para que os usuários retirem nas estações. Neste trabalho, estimamos o número de retiradas de bicicletas em cada estação do sistema Citi Bike, da cidade de Nova Iorque, para um determinado dia a partir de um modelo de predição de demanda que utiliza o algoritmo *K-Nearest-Neighbors (KNN)* e a similaridade das condições meteorológicas. Por fim, propomos um plano de implementação de um sistema para apoiar decisores a otimizar seu inventário utilizando o preditor em uma startup deste mercado.

Palavras-chave: Predição de demanda. mobilidade urbana. mobilidade inteligente. bike sharing. MSWK Regressor.

Demand Prediction for Station Based Bike Sharing Systems Using Meteorological Data Aggregation

ABSTRACT

Startups are young technological businesses. To be lean and competitive, it is important to stay essentialists and efficient in their activities. In station based bike sharing systems, we can verify efficient practices in balancing operations that maximize bike utilization. For this, it is necessary that the system offers a high availability of bikes for user demand. In this work, we estimate the number of daily bike withdraws in each station of Citi Bike, at New York City, using a prediction model based on the K-Nearest-Neighbors (KNN) algorithm and the similarity of weather conditions. Finally, we present a plan of a system implementation to support decision takers to optimize their inventory using our prediction model in a local startup.

Keywords:

.

LISTA DE ABREVIATURAS E SIGLAS

KNN	K-Nearest-Neighbors
MSWK	Meteorology Similarity Weighted KNN Regressor
GPRS	General Packet Radio Service
GPS	Global Positioning System
OSEMN	Obtain, Scrub, Explore, Model, Interpret
MAE	Mean Absolute Error
NOAA	National Oceanic and Atmospheric Administration

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A bicicleta, apesar de ser um modal de transporte antigo, vem se tornando cada vez mais comum nos sistemas de mobilidade urbana de grandes cidades. Por ser eficiente no transporte em trajetos de curta distância, também chamados de última/primeira milha, auxilia o transporte individual ao complementar os modais coletivos em viagens cotidianas de longa distância. Além disso, é um modal ativo que ajuda no combate ao sedentarismo da população. Esses trajetos curtos fazem parte da chamada *micromobilidade*, que, na última década, vêm recebendo a oferta de novos serviços pelo mundo todo. Com o crescimento acelerado, esses serviços estão alterando radicalmente a cultura e infraestrutura de transportes das cidades [11, 23].

Os sistemas implementados mais recentemente apresentam novas características como ofertar, além da bicicleta mecânica tradicional, outros modais como patinetes elétricos e bicicletas elétricas e mais flexibilidade de uso [25]. Conhecidos como *smart bikes* [24], esses sistemas operam com ou sem estações. Os sistemas sem estação (*dockless*) possibilitam iniciar e finalizar viagens em qualquer lugar dentro da área em que o serviço está disponível [9]. Já os sistemas com estações virtuais também não apresentam estações com baias físicas e permitem a devolução em estações definidas como pequenas áreas geográficas.

Em ambas as formas de operação, o controle das bicicletas é descentralizado utilizando um *cadeado inteligente*, possibilitando que a própria bicicleta controle sua trava de liberação por comunicação via GPRS e/ou *bluetooth* com o *smartphone* do usuário. Comparado com o modelo de estações físicas, este modelo de sistema possui como principal característica econômica ter um investimento significativamente menor, evitando a necessidade de custosas estações de controle. Entretanto, apresentam desafios como furtos, vandalismos, bloqueio de vias públicas e aumento nos custos de balanceamento, no caso *dockless*. Estes desafios tentam ser mitigados com o modelo de estações virtuais, como ocorre na operação da startup Loop Bike Sharing [1].

1.1 Loop Bike Sharing

A Loop é uma startup tecnológica que opera um sistema de *Smart Bike* em escala reduzida, com menos de 100 bicicletas e 30 estações virtuais, na cidade de Porto Alegre/RS [1]. Atualmente, seus processos de balanceamento são diários e a distribuição é

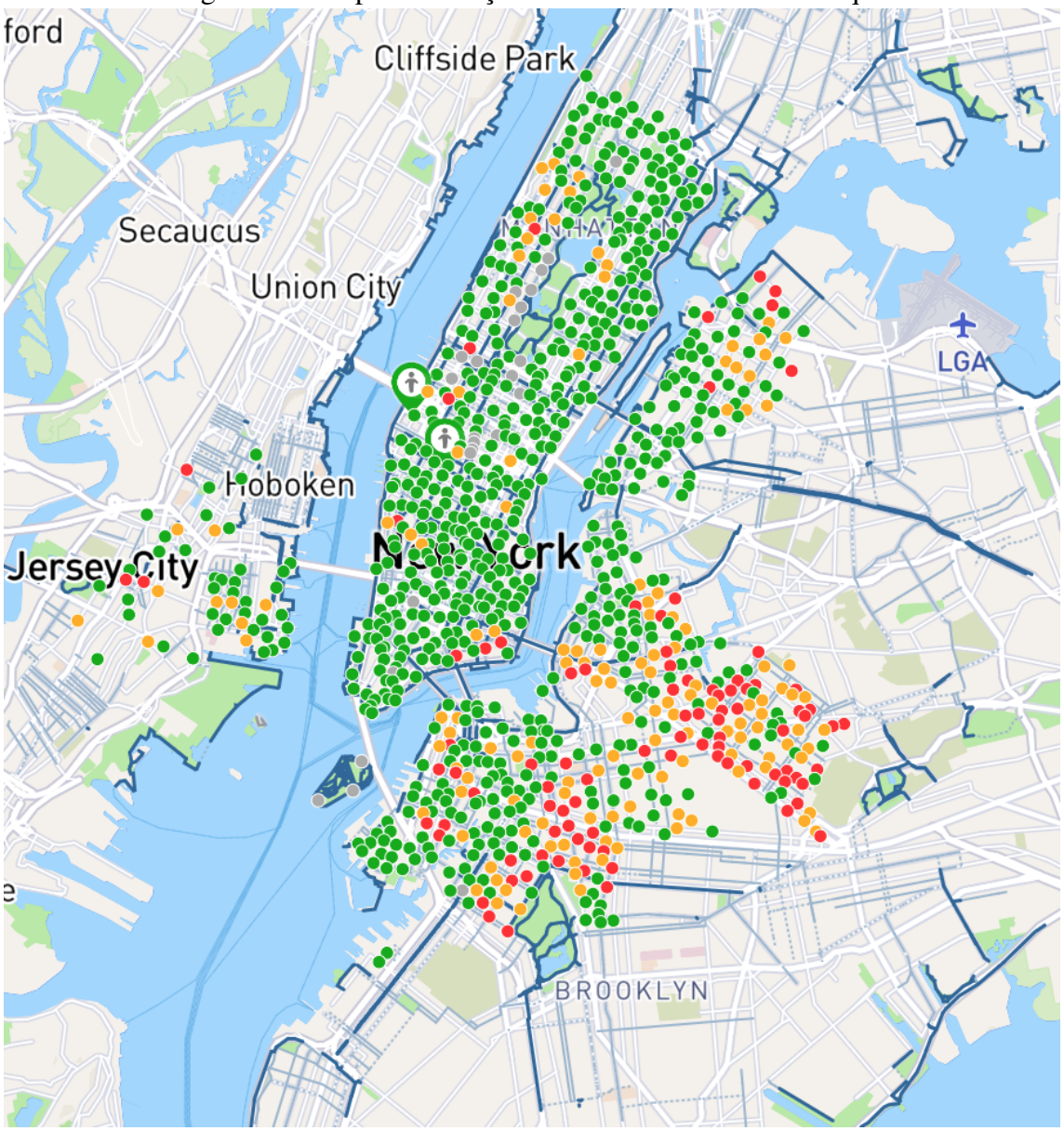
feita a partir da análise do histórico utilizando métodos simplificados com constantes.

Para o desenvolvimento do trabalho, utilizaremos os dados abertos disponíveis no sistema Citi Bike, operado pela empresa Motivate [2], por ser amplamente estudado [26, 23, 15, 14] e ter um longo histórico de utilização [3].

1.2 Bicicletas Compartilhadas em Nova Iorque

O sistema Citi Bike em operação na cidade de Nova Iorque, nos Estados Unidos da América, é do tipo que utilizada estações físicas e possui cerca de 12 mil bicicletas distribuídas em suas 750 estações presentes em espaços públicos [4]. Cada ponto na Figura 1.1 representa uma estação de retirada e devolução.

Figura 1.1: Mapa das estações no Citi Bike em Nova Iorque



A empresa operadora publica mensalmente relatórios com as viagens e o resumo da utilização do sistema [3]. Estes relatórios têm sido estudados em trabalhos desenvolvidos para problemas de predição de demanda e roteamento [26, 23, 15, 14].

Neste trabalho, desenvolvemos apenas o problema de predição de demanda para auxiliar na etapa de estimativa de target para os sistemas com balanceamento estático que possuem estações físicas ou virtuais e utilizamos os dados dos relatórios do sistema de Nova Iorque para treinar nosso modelo. Para a predição, utilizaremos uma adaptação do regressor *Meteorology Similarity Weighted K-Nearest-Neighbor* (MSWK) [15]. Por fim, este trabalho propõe um plano de implementação do modelo de predição na startup Loop Bike Sharing.

1.3 Balanceamento

A operação de balanceamento consiste em realocar as bicicletas disponíveis e recolher as que necessitam de reparos. Em geral, utilizam outros veículos maiores para carregá-las.

Existem dois tipos de balanceamento: *estático* e *dinâmico* [10]. O balanceamento estático considera que as bicicletas não saem das estações durante todo o período de operação do balanceamento. O balanceamento dinâmico considera que, enquanto ocorre o balanceamento, os usuários continuam retirando, utilizando e devolvendo as bicicletas. Em sistemas reais, não é viável parar a operação para o balanceamento, então é comum que este processo ocorra a noite e ao início do dia, nos horários em que o sistema tem baixíssima utilização.

Ambos os tipos de balanceamento se dividem nas etapas a seguir.

1.3.1 Estimativa de Target

O *target* define quantas bicicletas devem ficar disponíveis na estação considerando as bicicletas em bom estado no sistema, a capacidade física da estação e a expectativa de retiradas e devoluções para a estação no período entre cada balanceamento.

A estimativa de *target* de cada estação é necessária para o melhor balanceamento do sistema, buscando otimizar a disponibilidade das bicicletas para a população, resultando na melhora da qualidade do serviço [23].

Após estimar o target, o operador pode planejar a distribuição para cada estação, as rotas que serão executadas e quantos veículos de carga serão utilizados.

1.3.2 Roteamento

Para planejar a distribuição das bicicletas e as rotas a serem feitas pelos veículos de carga, os operadores precisam de um target definido e da lista de bicicletas que serão recolhidas para serem manutenidas.

As rotas dos veículos que distribuem as bicicletas são feitas visando minimizar os custos de hora-homem e de combustível, bem como o tamanho da frota de veículos de carga [21].

A organização do trabalho segue com o capítulo 2 que relaciona os trabalhos anteriores e as abordagens mais comuns na literatura; capítulo 3 define o problema de predição de demanda em um sistema de bicicletas compartilhadas com estações; capítulo 4 explica nossa metodologia e os conceitos utilizados no modelo; capítulo 5 explica as fontes e manipulações dos dados; a capítulo 6 analisa os resultados obtidos na análise exploratória e nos testes do modelo; capítulo 7 contém as conclusões do trabalho e um plano de implementação na startup Loop Bike Sharing.

2 TRABALHOS ANTERIORES

A literatura de predição de demanda é extensa e muito rica, em geral, focando em problemas que maximizam lucratividade e qualidade de serviços, na distribuição de ativos. Porém, as pesquisas de sistemas de bicicletas compartilhadas ainda é recente e tem evoluído rapidamente com o aumento do impacto deste tipo de serviço no cotidiano das grandes cidades. Analisando o conhecimento produzido nos últimos anos, vemos algumas principais abordagens.

2.1 Constantes

É comum o uso de simplificações na predição da demanda de uma estação utilizando números conhecidos, como a definição de constantes como metade das vagas em uma estação física [18, 19, 20]. Nestes métodos, o objetivos está apenas em estimar um *target* para estações em um sistema teórico e sem histórico de viagens no mundo real. Em geral, são trabalhos que focam em soluções para construção de rotas de balanceamento com veículo único ou múltiplos veículos.

2.2 Métodos estatísticos

Em outros casos, o *target* é uma média histórica da estação [13]; uma faixa de valores baseados na satisfação dos usuários [21]; uma expectativa de demanda usando *processo de Poisson* [8]; ou *métodos estocásticos* [22]. São sistemas que levam em conta o histórico de viagens de usuários reais reconhecendo as peculiaridades na demanda de cada estação. Contudo, estes métodos não levam em conta variáveis de influência direta ou indireta na demanda, deixando de apresentar as características de sazonalidade que são observadas no mundo real.

2.3 Agregação de dados

Propostas utilizando outros fatores de influência, diretamente ou indiretamente relacionados aos serviços, trazem uma nova perspectiva para problemas de adaptação do sistema à demanda no mundo real. A agregação de dados meteorológicos ao histórico é

promissora na melhora de acurácia da predição quando existe histórico de uso [15]. Já em casos onde não há histórico, podemos utilizar a agregação de dados de tráfego e *pontos de interesse* da demanda na cidade para auxiliar novos projetos de expansão de sistemas para novas áreas [14].

3 PROBLEMA

Nesta seção, definimos formalmente as estruturas e elementos básicos utilizados no trabalho. Após, definiremos o problema de predição de demanda de bicicletas para as estações da rede.

3.1 Rede de Tráfego

Uma rede de tráfego é definida por um grafo $G = (S, E)$ onde os nodos $s \in S$ representam estações e as arestas $e \in E$ compostas por $e_{ij} = (s_i, s_j)$ representam o tráfego vindo da estação s_i para s_j.

Construímos a rede de tráfego a partir de um conjunto histórico de viagens TR, onde $tr \in TR$ e $tr = (s_a, s_b, t_a, t_b)$ representa uma viagem vindo de s_a e indo para s_b, iniciando em t_a e terminando em t_b, representados em minutos e não incluindo viagens em que $t_b - t_a < 1$ minuto.

3.2 Demanda

Cada estação possui uma demanda diária de retiradas de bicicletas $s_i.pd(t)$ definida pelo total de viagens em TR, tal que $tr = (s_i, *, t, *)$.

Na seção 5, detalhamos como $s_i.pd$ precisou ser normalizado para que o erro de predição em estações de menor demanda não distorcessem a avaliação dos resultados.

3.3 Predição de Demanda

Problemas de predição podem ser divididos em dois tipos: classificação e regressão. Os problemas de *classificação* têm o objetivo de predizer classes discretas. Já os problemas de *regressão* predizem valores contínuo.

Os valores de demanda de viagens das estações são valores contínuos e nosso preditor é um regressor que, dado um histórico de viagens de bicicletas diárias TR_t e um conjunto de dados meteorológicos W_t, prediz a demanda diária de retiradas de bicicletas $s_i.pd(t)$ para cada estação em S.

4 METODOLOGIA

Como este é um trabalho que adapta o método do regressor MSWK [15], aplicamos sobre a mesma base de dados do Citi Bike, aprofundando a busca de correlações entre dados meteorológicos e demanda e adaptando para o caso da predição de demanda com frequência diária.

Para a sistematização do trabalho, seguimos a organização proposta pelo *framework OSEMN* [16]. Nele é proposto o fluxo de trabalho da Figura 4.1, onde obtemos os dados, limpamos, exploramos suas características, construímos o modelo de predição e, por fim, interpretamos os resultados.

Figura 4.1: Fluxo de trabalho

Os experimentos que praticamos visam análises exploratórias que identifiquem as características da demanda do sistema, além de parametrizarem nosso regressor conforme os atributos meteorológicos, o histórico de viagens e o aprendizado do algoritmo.

4.1 Análise Exploratória

Cada sistema de mobilidade apresenta diferentes características pois recebe influência de diversos fatores do ambiente, como condições meteorológicas, de infraestrutura de transporte e culturais. Por isso, buscamos identificar as características do sistema observado a partir do histórico de viagens e de dados meteorológicos.

4.1.1 Características do Citi Bike

As principais características que buscamos são: a de *sazonalidade de demanda* durante o ano, que deve acompanhar as mudanças de estações do ano e suas características meteorológicas; e a diferença de uso entre *dias de semana* e *finais de semana*, refletindo fatores culturais que levam as pessoas ao uso para transporte ou para lazer.

Existem diversos outros fatores que podem ser analisados, porém entendemos que

estes mostram as maiores variações na demanda das bicicletas. Eventos pontuais como feriados, comemorações ou competições esportivas não foram levados em conta.

4.1.2 Seleção de Atributos Meteorológicos

Calculamos as correlações dos atributos meteorológicos com a demanda para definir que atributos são mais relevantes para modelagem de bons preditores. Com os atributos de maior correlação direta ou inversa calculamos variáveis de similaridade meteorológicas entre os dias, utilizadas no regressor.

4.2 MSWK Regressor

Implementaremos o método *Meteorology Similarity Weighted KNN Regressor (MSWK)* [15] que agrega dados meteorológicos ao algoritmo K-Nearest Neighbors (KNN) [7] e utiliza funções Gaussianas para determinar similaridades meteorológicas. Nossa implementação é adaptada para o problema deste trabalho, em que os períodos são diários e não por hora do dia.

4.2.1 K-Nearest Neighbors

Em problemas de regressão, o algoritmo gera um ponto no espaço de atributos utilizando a média dos atributos das K observações mais similares [7]. Em nossos experimentos, o ponto e o espaço possuem duas dimensões que se referem às duas variáveis de distância. Como medida de distância, usaremos a similaridade dos atributos meteorológicos selecionados na etapa de análise exploratória com um peso α que será aprendido pelo MSWK Regressor. Os melhores K encontrados são definidos a partir da maior precisão de predições para dias de semana e para finais de semana.

4.2.2 Similaridade Meteorológica

Os atributos meteorológicos possuem diferentes correlações com a demanda. Por isso utilizamos uma *Gaussian Kernel function* para avaliar a similaridade entre atributos

meteorológicos entre dois dias D_p e D_q no algoritmo KNN.

Agrupamos os atributos conforme a sensibilidade da demanda. Assim, os atributos aos quais a demanda é mais sensível foram avaliados separadamente utilizando a seguinte fórmula:

$$\lambda_1(X_{D_p}, X_{D_q}) = \frac{1}{2\pi\sigma} \mathrm{e}^{\frac{(X_{D_p} - X_{D_q})^2}{\sigma^2}} \tag{4.1}$$

Já os atributos dos quais a demanda tem sensibilidade semelhante foram agrupados seguindo uma *n-D Gaussian Kernel*, onde n é a quantidade de atributos:

$$\lambda_2(Y_{D_p}, Y_{D_q}) = \frac{1}{2\pi\sigma} \mathrm{e}^{-}(\sum_{i=1}^{n} \frac{(y_{i_{D_p}} - y_{i_{D_q}})^2}{\sigma^2}) \tag{4.2}$$

Onde Y é o conjunto dos valores de atributos agrupados por sensibilidade da demanda. Para simplificar, definimos todos os $\sigma = 0$.

4.2.3 Similaridade Geral

A função de similaridade resultante é formada a partir de uma combinação linear das similaridades meteorológicas apresentadas anteriormente.

$$M(D_p, D_q, \alpha) = \delta_w(D_p, D_q) \sum_{i=1}^{2} \alpha_i \lambda_i \tag{4.3}$$

Onde α é um conjunto de parâmetros aprendidos pelo regressor e δ_w indica se os dias D_p e D_q são dias de semana ou finais de semana, caso sejam do mesmo tipo então $\delta_w(D_p, D_q) = 1$, caso contrário, $\delta_w(D_p, D_q) = 0$, zerando toda a função de similaridade geral. Feriados serão considerados dias comuns em nossos experimentos.

4.2.4 Regressão

A predição de demanda $\hat{s}_i.pd$ é calculada a partir do dia que se deseja predizer demanda D_q e do conjunto de parâmetros aprendidos α.

$$\hat{s}_i.pd(D_q, \alpha) = \frac{\sum_{p=1}^{K} M(D_p, D_q, \alpha) s_i.pd(D_p)}{\sum_{p=1}^{K} M(D_p, D_q, \alpha)} \tag{4.4}$$

Onde K é o total de dias mais semelhantes a D_q e $s_i.pd(D_p)$ é a demanda observada para cada dia D_p.

O aprendizado dos parâmetros α é feito minimizando o *Mean Absolute Error (MAE)* da predição $\hat{s}_i.pd$, sendo $\alpha*$ o conjunto de menor MAE encontrado por *força bruta*.

$$\alpha* = \arg\min_{\alpha} \frac{1}{N} \sum_{i=1}^{N} |\hat{s}_i.pd(D_q, \alpha) - s_i.pd(D_q)| \tag{4.5}$$

Onde N é o total de estações do sistema.

4.3 Avaliação

Para avaliar a precisão do método de predição de demanda usamos a métrica *Mean Absolute Error* (MAE), quanto menor o erro absoluto, maior é a precisão.

4.3.1 Mean Absolute Error

O MAE nos mostra a média do erro absoluto do resultado predito para o observado. Definimos esta métrica como:

$$MAE = \frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{i=1}^{m} |\hat{s}_i.pd(t) - s_i.pd(t)|}{m} \tag{4.6}$$

Onde T é o total de dias, m é o total de estações e $\hat{s}_i.pd(t)$ e $s_i.pd(t)$ são a predição de demanda para a estação e a demanda observada no dia t, respectivamente.

4.3.2 Validação Cruzada

Estratégias de validação cruzada são importantes para evitar casos de *overfitting*, quando o modelo aprendido funciona bem apenas para o conjunto em que foi treinado. Existem diversos métodos de fazer a validação cruzada, que, em geral, consistem em separar as amostrar entre um conjunto de treinamento e um conjunto de teste.

Para validarmos os resultados, dividimos as observações em dois conjuntos, em 70/30. Onde 70% das observações foram utilizadas para treinamento do regressor e as

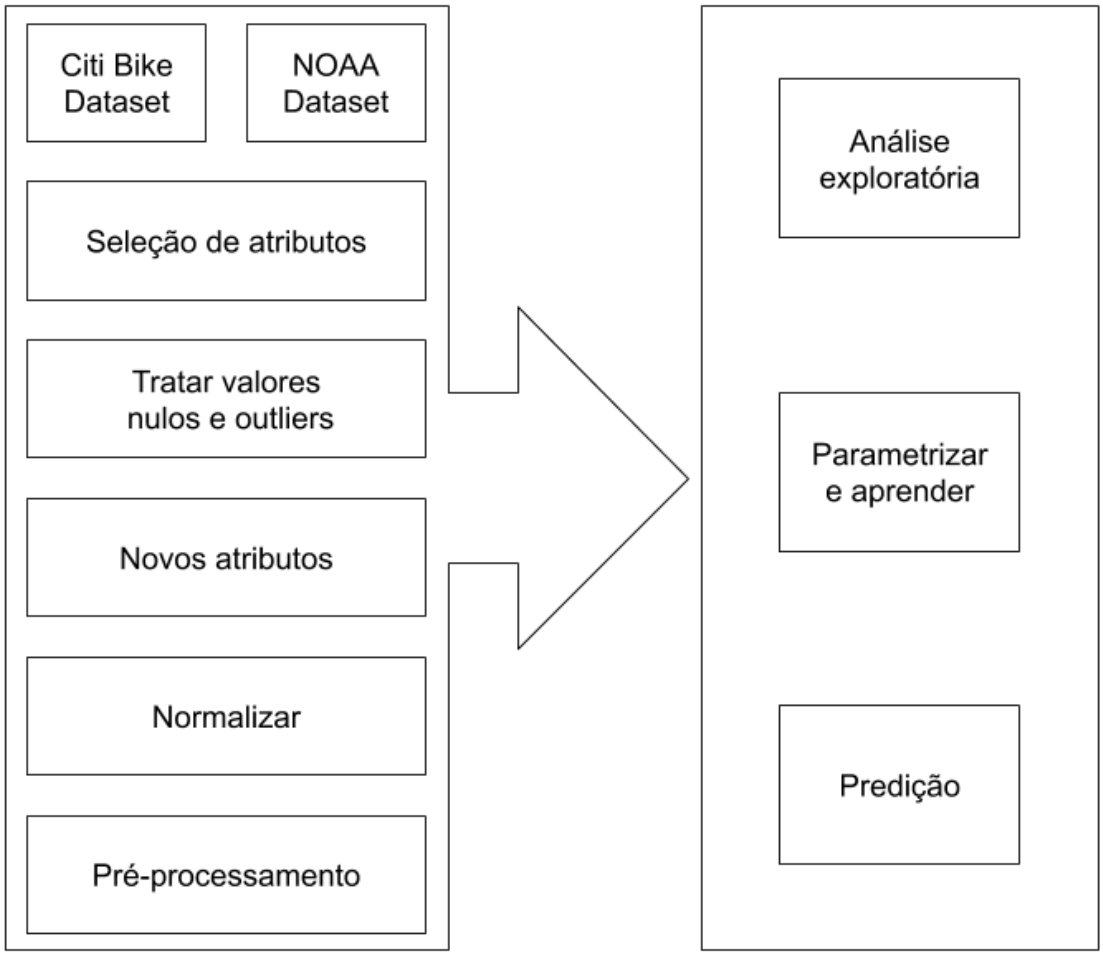
demais 30% para testar o modelo aprendido.

5 OBTENÇÃO E LIMPEZA DE DADOS

As diversas fontes de dados possuem peculiaridades em suas representações, como dados faltantes ou nulos, presença de outliers e muitos atributos que não iremos utilizar nos experimentos. Assim, antes de qualquer análise ou predição, após a obtenção dos dados, precisamos fazer alguns tratamentos e pré-processamentos.

A Figura 5.1 apresenta os passos que executamos antes das fases de exploração e construção do modelo de predição.

Figura 5.1: Fluxo de trabalho detalhado



5.1 Fontes de Dados

Contamos com duas fontes de dados neste trabalho: os dados abertos pela *Motivate International Inc*, empresa que opera o sistema Citi Bike; e dados meteorológicos das estações de medição na cidade de Nova Iorque, disponibilizados pela *National Oceanic*

and Atmospheric Administration (NOAA).

5.1.1 Citi Bike

Os dados do sistema Citi Bike são divididos em relatórios mensais e estão disponíveis em formato $.csv$. Para os experimentos extraímos os registros mensais e agrupamos em um arquivo único para o período entre 01/07/2013 e 28/05/2015 [3]. Selecionamos os atributos relevantes para os experimentos e agregamos os novos atributos que representam a duração, o dia da semana e se os dias são finais de semana, para serem analisados na etapa exploratória. Assim, os registros de viagens ficaram com os seguintes atributos:

- hora de início;
- data de início;
- duração, em minutos;
- identificador da estação de retirada
- identificador da estação de devolução
- dia da semana;
- final de semana, em booleano;

5.1.2 Meteorológicos

As observações do NOAA são registradas por dia para cada estação meteorológicas distribuída pela cidade. Os dados são solicitados pelo site da NOAA [5] e foram disponibilizados em formato $.csv$ e nenhuma sanitização foi necessária. Os atributos meteorológicos de cada dia são gerados a partir das médias diárias das medições das estações.

O conjunto de atributos meteorológicos gerados a partir das medições foram os seguintes:

- temperatura máxima, em °C;
- temperatura mínima, em °C;
- temperatura média, em °C;
- volume de chuva, em mm;
- velocidade média do vento, em m/s;

Os registros meteorológicos originais têm muitos dados nulos entre as estações de medição. Para tratar, removemos as estações sem medição, assim as médias foram calculadas somente com valor não nulos.

5.2 Normalização

O processo de normalização coloca os valores de um atributo em uma mesma faixa de valores. Isso faz com que atributos com valores originalmente em escalas diferentes estejam, agora, em mesma escala e possibilitem comparação.

Para os cálculos de similaridade meteorológica serem comparáveis na construção do modelo, os valores diários de total de viagens e dados meteorológicos foram normalizados, recebendo valores contínuos entre 0 e 1.

Para normalizar os dados, utilizamos o método *Min-Max*, que aplica o seguinte cálculo:

$$X_{normalizado} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{5.1}$$

Onde X é o valor a ser normalizado e X_{min} e X_{max} são o menor e o maior valor observado para algum X, respectivamente.

5.3 Pré-processamento

Ao juntar os arquivos de histórico mensal de viagens, o arquivo resultante tinha um tamanho de 1.2GB. Arquivos grandes como esse fazem com que cada etapa de processamento dure cerca de uma hora em um MacBook Intel Core i5 2,8GHz com 8GB DDR3.

Assim, criamos arquivos de *checkpoint*, em formato $.h5$, com o resultado de cada etapa de processamento do modelo, facilitando o trabalho de análise exploratória e construção do modelo.

6 RESULTADOS

Os experimentos a seguir foram executados no computador citado anteriormente na plataforma MacOS Mojave utilizando Jupyter Notebook e a linguagem Python 3 com o auxílio das bibliotecas Pandas e Numpy para cálculos de correlação, uso de estruturas de dados e suas manipulações.

Abaixo, a análise exploratória mostra as características da demanda, após construímos e avaliamos o modelo de predição.

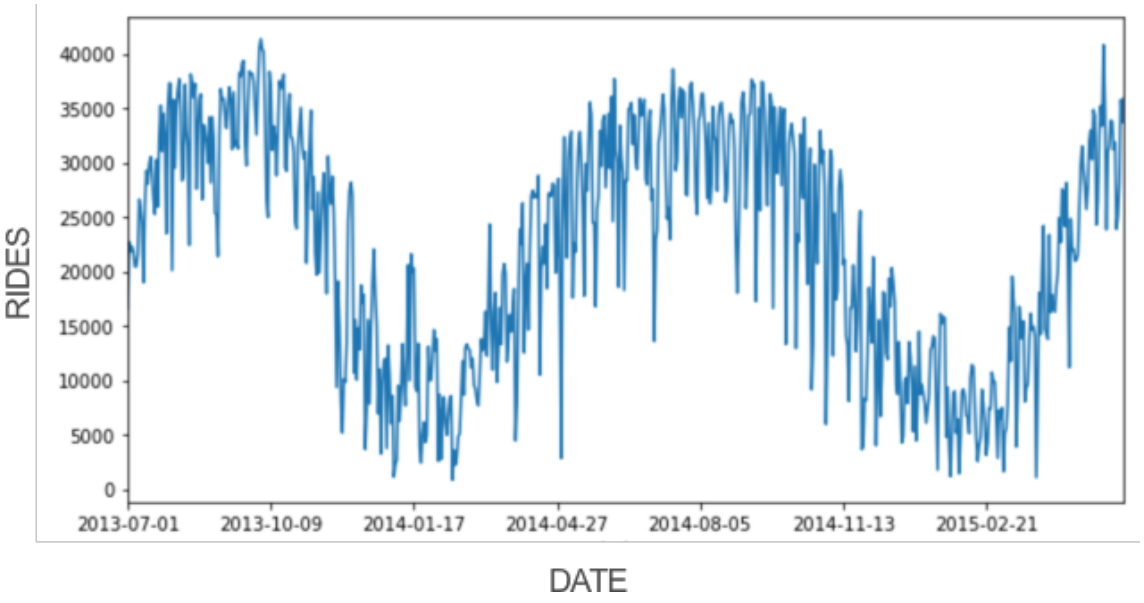
6.1 Análise Exploratória

O sistema Citi Bike apresenta características de sazonalidade no ano e diferentes demandas nas estações entre dias de semana e finais de semana. A seguir, mostramos como tais características foram observadas a partir da análise da demanda.

6.1.1 Sazonalidade

A escolha de agregação dos dados meteorológicos se baseia no efeito da sazonalidade de demanda em sistemas de bicicletas compartilhadas. Efeito que se evidenciou no Citi Bike durante período das amostras.

Figura 6.1: Sazonalidade durante o ano



A sazonalidade pode ser explicada por sua forte correlação com a temperatura

média do dia, mostrada na análise de atributos meteorológicos. Assim, como o clima da cidade é marcado por uma grande amplitude térmica durante o ano, podemos verificar grande amplitude nas demandas diárias durante o ano, com seus mínimos e máximos ocorrendo durante os períodos de inverno e verão, respectivamente.

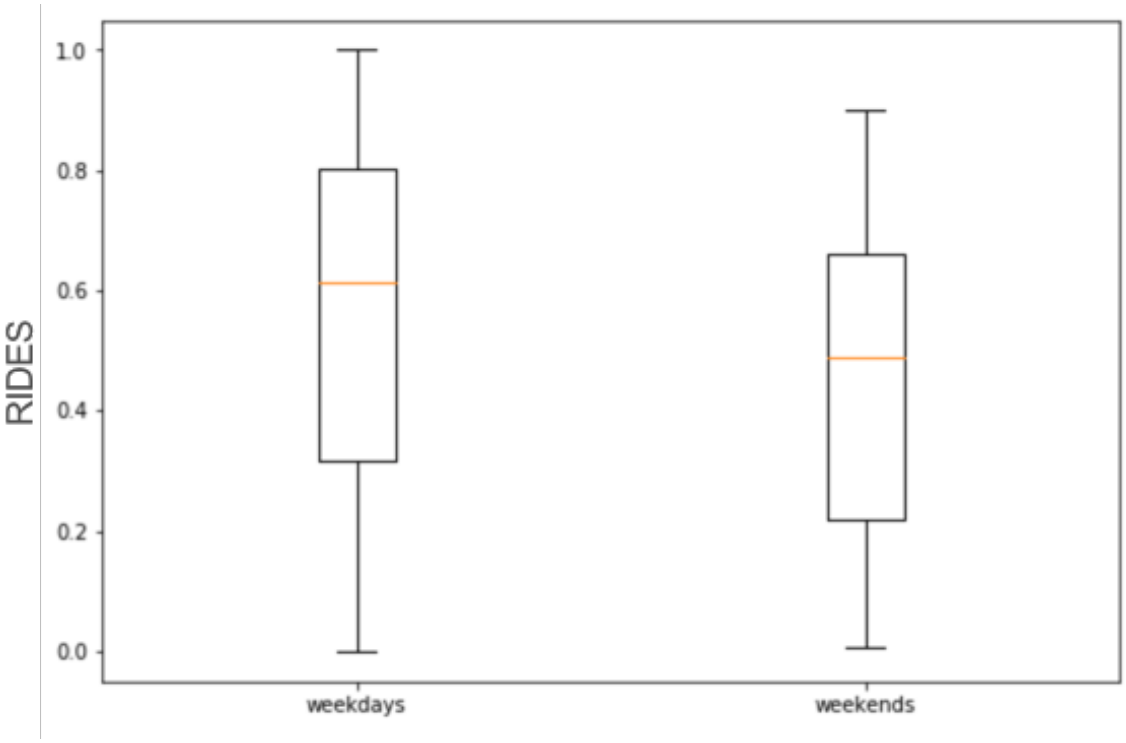
6.1.2 Dias Úteis e Finais de Semana

Os comportamentos dos usuários em dias úteis e finais de semana e feriados diferem bastante. As demandas em dias úteis sendo mais focadas em mobilidade, e nos finais de semana e feriados predomina o uso para lazer.

Em geral, esses comportamentos são caracterizados pela duração de viagem, sendo possível identificar trajetos de mobilidade e de lazer pelo tempo que os usuários demoram. Viagens de mobilidade se caracterizam por terem duração similar à estimada para o melhor trajeto estimado. Já as viagens de lazer têm duração significativamente maior comparadas às de mobilidade.

Na Figura 6.2, vemos que a demanda em dias de semana possui mais viagens e uma máxima maior se comparada aos finais de semana.

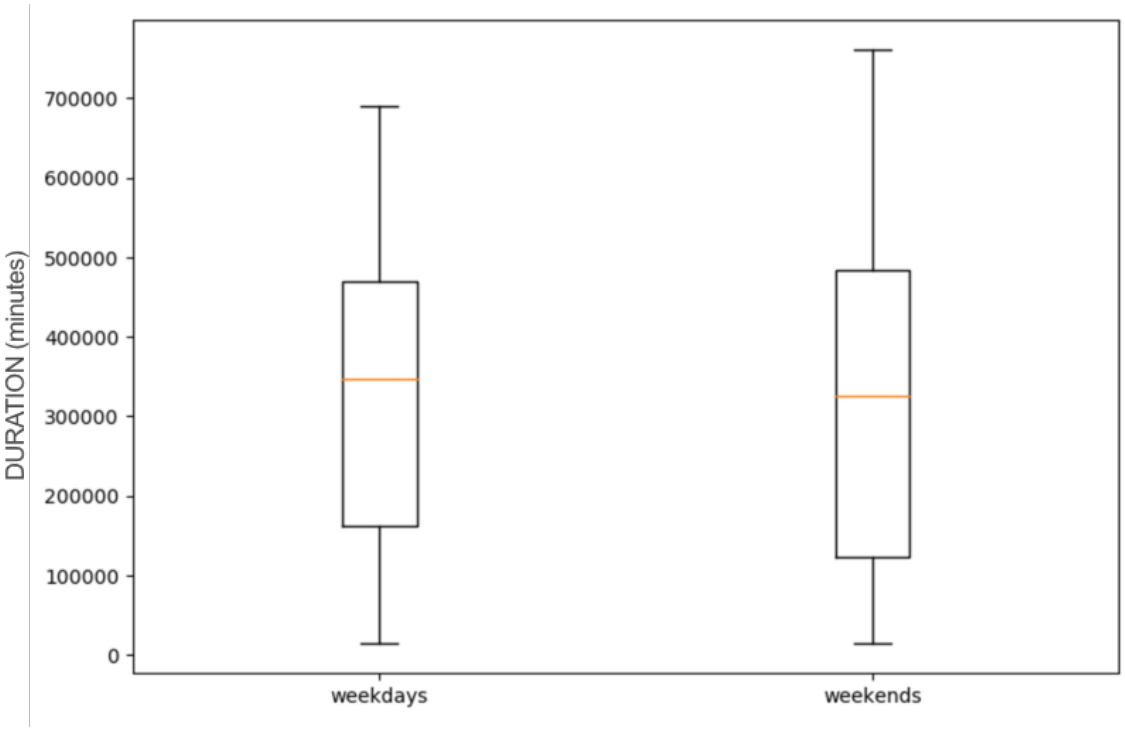
Figura 6.2: Comparação entre demandas em dias de semana e finais de semana



Mesmo com um total de viagens inferior em finais de semana, ao analisarmos o

somatório de tempo de viagens podemos observar um possuem um máximo maior e uma distribuição mais espaça, caracterizando o aumento de dias com viagens mais longas que os dias de semana.

Figura 6.3: Comparação entre duração total de viagens em dias de semana e finais de semana



6.2 Predição

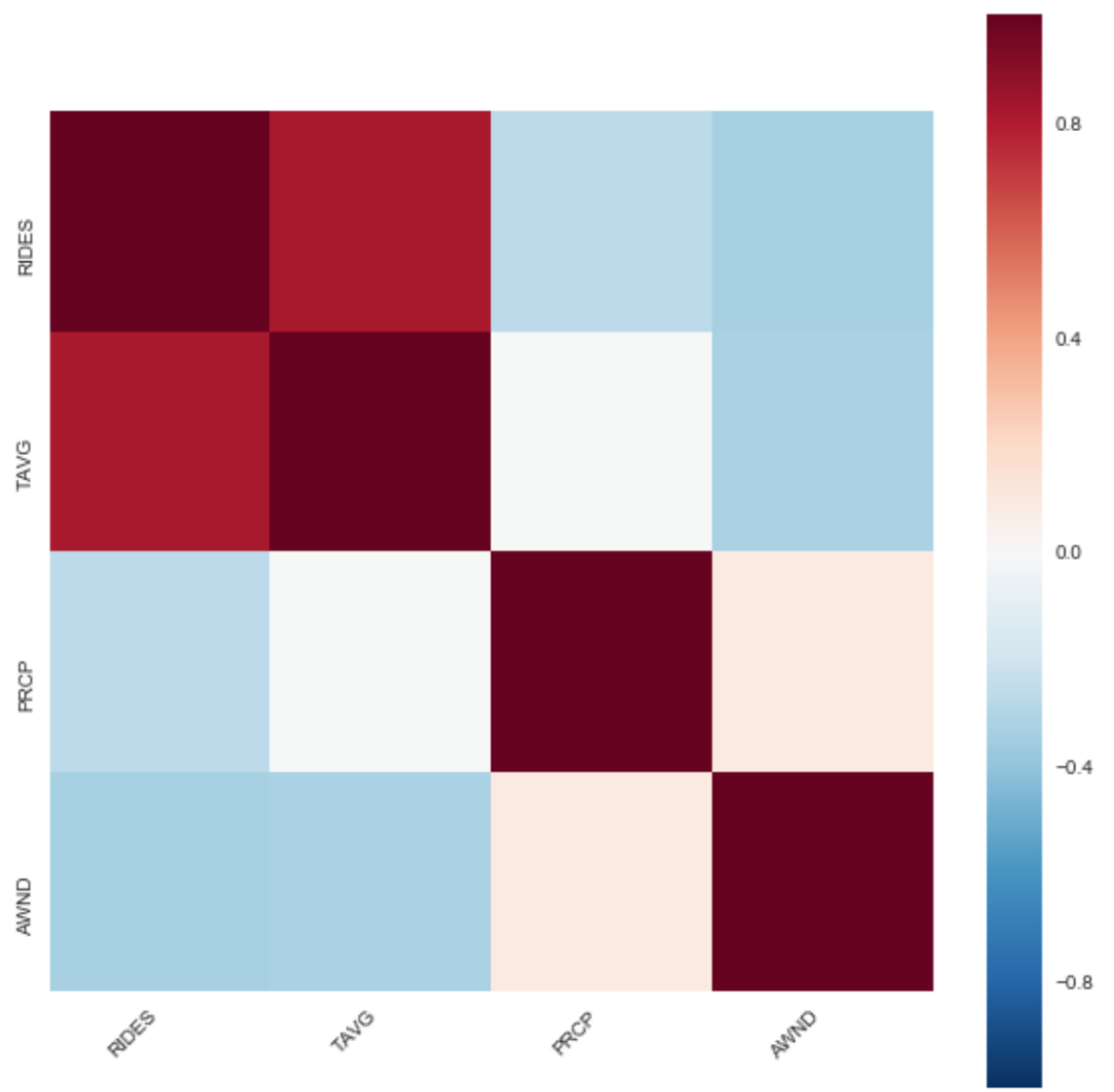
Para a construção do modelo de predição, verificamos a correlação dos atributos meteorológicos para gerar os atributos de similaridade λ. Após gerados, treinamos o nosso regressor para aprender os pesos α e a quantidade de dias mais similaridades K que resultaram as maiores precisões para dias de semana e finais de semana para todo o período.

6.2.1 Atributos Meteorológicos

Utilizamos a *correlação de Pearson* para verificar a existência de uma relação de linearidade entre os atributos meteorológicos e a demanda. Os resultados nos mostram que as demandas diárias têm forte correlação positiva com a temperatura média e fraca correlação negativa com o volume de precipitação e velocidade média do vento, conforme

a *matriz de correlação* abaixo.

Figura 6.4: Matriz de correlação



Nos gráficos abaixo, podemos observar como os atributos temperatura média, chuva e velocidade do vento influenciam na variação da demanda diária.

Figura 6.5: Demanda x temperatura média

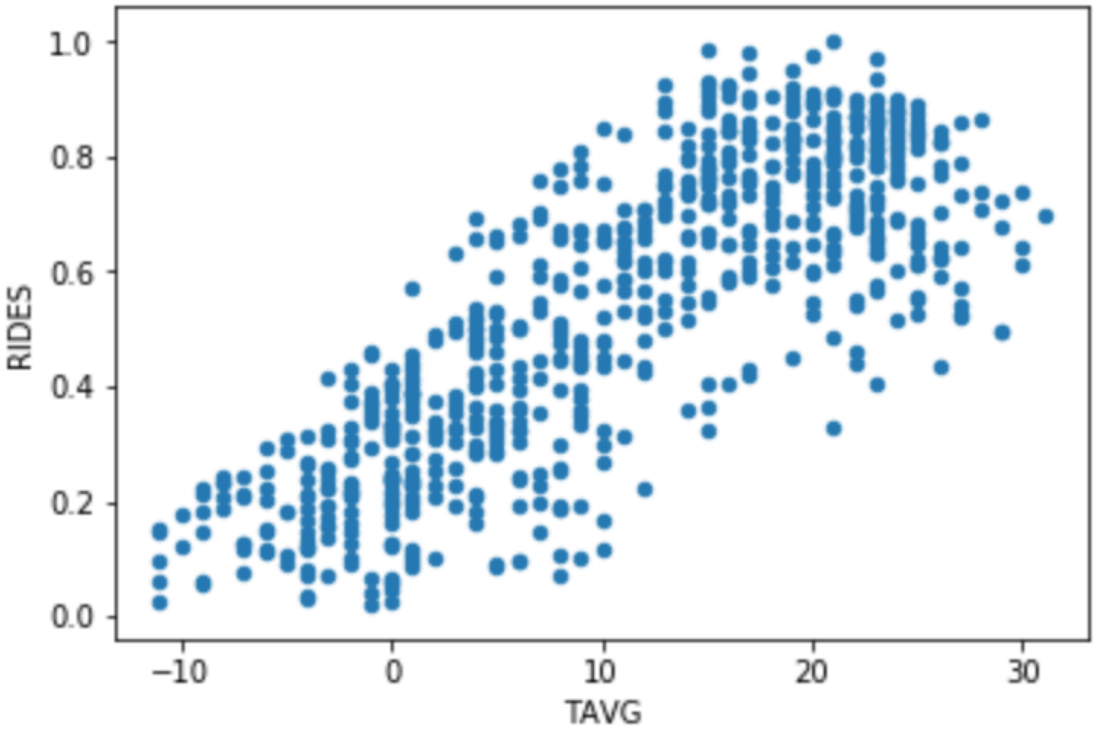


Figura 6.6: Demanda x volume de chuva

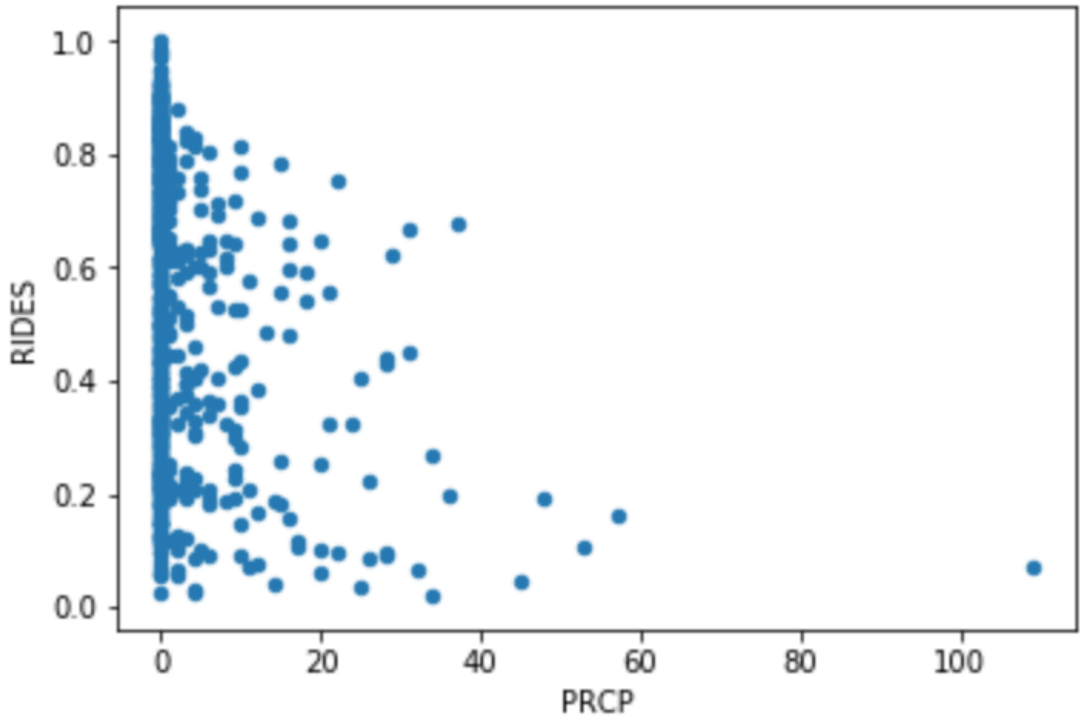
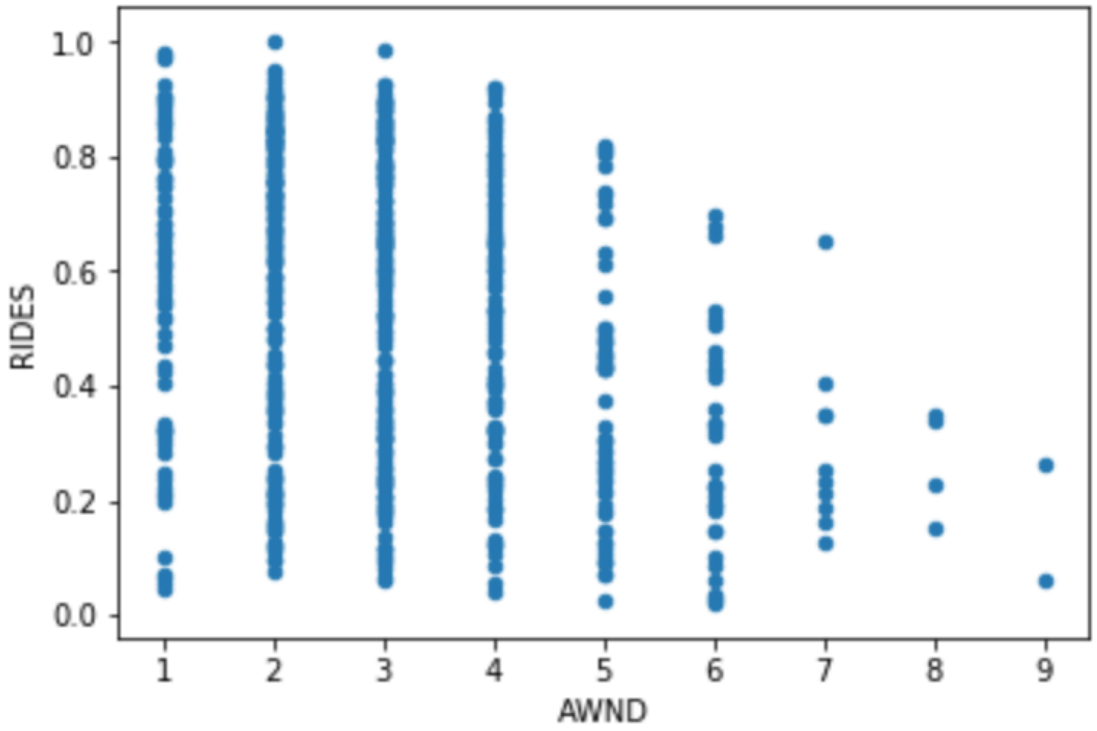


Figura 6.7: Demanda x velocidade média do vento



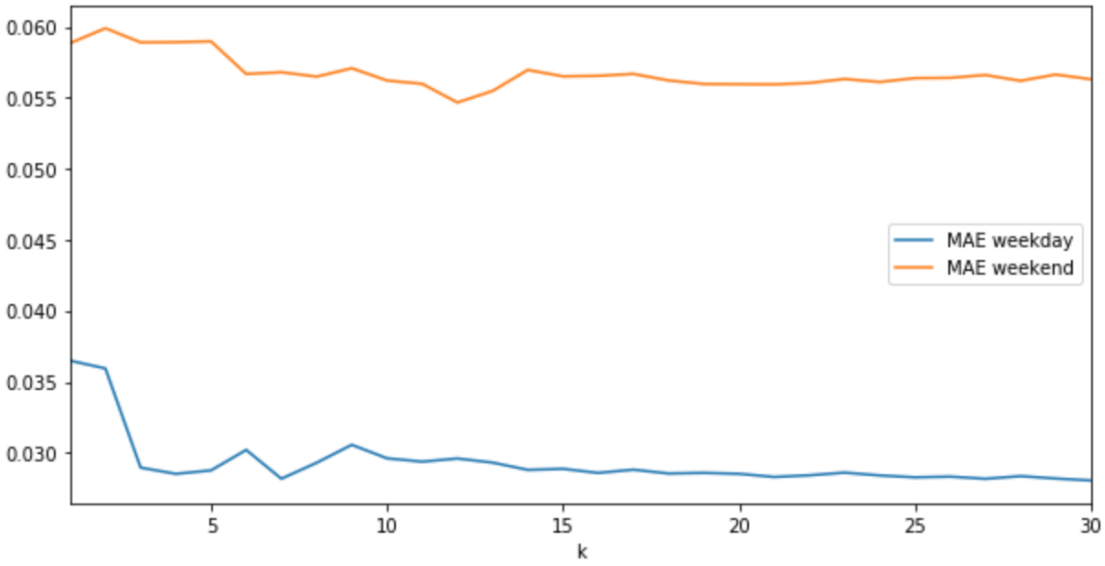
Então, por ter forte correlação com a demanda, transformamos a temperatura média em um parâmetro de similaridade λ_{tavg} utilizando a função gaussiana com uma dimensão. Os demais atributos, velocidade do vento e chuva, têm correlação fraca e negativa com a demanda, por isso serão agrupados em apenas um parâmetro de similaridade $\lambda_{awnd-prcp}$, utilizando uma gaussiana de 2 dimensões.

6.2.2 Parametrização

O aprendizado dos pesos α foi feito por *força bruta*, gerando 10 tuplas aleatórias com números contínuos entre 0 e 10 e medindo a utilidade a partir do menor MAE resultante. Não houve convergências em tuplas específicas, logo, a cada bateria de treinos, o modelo aprendeu novos α, mantendo um MAE semelhante.

Os parâmetros k (dias-mais-semelhantes) foram aprendidos, separadamente, para dias de semana e finais de semana do conjunto de testes. O critério para escolha dos melhores k foi o menor MAE.

Figura 6.8: K dias mais similares



O gráfico 6.8 mostra os parâmetros k que resultaram na melhor precisão. Ao utilizar $k = 30$, obtivemos $MAE = 0.0280$ para dias de semana; e $k = 12$, gerou $MAE = 0.0546$ para finais de semana, em uma escala de demanda normalizada entre [0,1].

7 CONCLUSÃO

A predição de demanda no sistema estudado mostrou que as condições meteorológicas e as características culturais, como os hábitos de uso da população, podem ser usadas de forma satisfatória para predição de demanda. Especialmente, o atributo de temperatura média, possui forte correlação com a demanda. Assim, utilizar esses atributos ambientais, do mundo real, contribuem para preditores em sistemas reais de bicicletas compartilhadas.

Apesar de dias de semana terem tido o melhor resultado com $k = 30$, acreditamos que a utilização de um k entre 14 e 18 seja mais proveitoso por reduzir a quantidade de dados processados, diminuindo o custo computacional.

7.1 Plano de Implementação

A startup *Loop Bike Sharing* pretende implementar um sistema que auxilie os operadores na tomada de decisão do target para cada estação em seus processos de rebalanceamento diários.

Na primeira etapa, o sistema iniciará com a implementação do método desenvolvido neste trabalho e irá comparar com seus métodos atuais que utilizam targets constantes. A avaliação será feita por métricas de qualidade do serviço: satisfação dos usuários e sua utilização. Na segunda etapa, será abordada a predição de sistemas de bicicletas compartilhadas *dockless*.

Como o foco de operação da empresa está no sul do Brasil, esperamos que as variações climáticas e culturais de cada cidade afetem de forma diferente as demandas. Outros fatores que devem ser entendidos pela empresa são os impactos de sistemas multimodais, adicionando bicicletas elétricas e patinetes, na demanda em sistemas. Da mesma forma, os efeitos de escala devem alterar o comportamento dos usuários.

8 TRABALHOS FUTUROS

O método utilizado neste trabalho agrega o histórico de viagens de um sistema que já estão em operação. No mundo real, nem sempre temos um histórico à disposição. Por isso, é importante explorar métodos que se baseiam em agregação de outros dados. Um trabalho recente propõe um método de predição da demanda de bicicletas para o planejamento da expansão do sistema Citi Bike para áreas que ainda não possuem um sistema de bicicletas, logo sem histórico [14]. Neste caso, outro dado de trânsito foi agregado, o tráfego de táxis nas áreas estudadas, e chegaram a bons resultados comparando a demanda predita com a demanda efetivamente observada.

Com o aumento do volume e diversidade de dados agregados, algoritmos baseados em *random forrest* e *redes neurais* se mostram promissores em sistemas reais com estações físicas, com predições mais precisas [17]. Outros efeitos como os picos de demanda, que causam problemas de falta de bicicletas, também apresentam resultados promissores quando tratados com técnicas baseadas em *redes neurais* [12].

Em sistemas sem estações, existem novos desafios para a predição de demanda que exigem novas abordagens. Nestes, a utilização de variáveis espaço-temporais do histórico de uso e de *long short-term memory network* se mostra promissor [6].

Por fim, as diferenças na escala e ambientais dos sistemas de bicicletas compartilhadas devem ser mais estudados para testar seus impactos na demanda e no acurácia do modelo *MSWK regressor*.

BIBLIOGRAFIA

[1] URL: <https://www.voudeloop.com>. (acessado em 21.04.2019).

[2] URL: <https://www.citibikenyc.com>. (acessado em 21.11.2019).

[3] URL: <https://www.citibikenyc.com/system-data/operating-reports>. (acessado em 21.04.2019).

[4] URL: <https://member.citibikenyc.com/map/>. (acessado em 21.11.2019).

[5] URL: <https://www.nws.noaa.gov/climate.php/>. (acessado em 21.04.2019).

[6] Yi Ai et al. "A deep learning approach on short-term spatiotemporal distribution forecasting of dockless bike-sharing system". Em: *Neural Computing and Applications* 31.5 (2019), pp. 1665–1677.

[7] Naomi S Altman. "An introduction to kernel and nearest-neighbor nonparametric regression". Em: *The American Statistician* 46.3 (1992), pp. 175–185.

[8] Ramon Alvarez-Valdes et al. "Optimizing the level of service quality of a bike-sharing system". Em: *Omega* 62 (2016), pp. 163–175. ISSN: 0305-0483. DOI: <https://doi.org/10.1016/j.omega.2015.09.007>. URL: <http://www.sciencedirect.com/science/article/pii/S0305048315002005>.

[9] Leonardo Caggiani et al. "A modeling framework for the dynamic management of free-floating bike-sharing systems". Em: *Transportation Research Part C: Emerging Technologies* 87 (2018), pp. 159–182. ISSN: 0968-090X. DOI: <https://doi.org/10.1016/j.trc.2018.01.001>. URL: <http://www.sciencedirect.com/science/article/pii/S0968090X18300020>.

[10] Daniel Chemla, Frédéric Meunier e Roberto Wolfler Calvo. "Bike sharing systems: Solving the static rebalancing problem". Em: *Discrete Optimization* 10.2 (2013), pp. 120–146.

[11] Paul DeMaio. "Bike-sharing: History, Impacts, Models of Provision, and Future". Em: *Journal of Public Transportation* 12 (dez. de 2009). DOI: <10.5038/2375-0901.12.4.3>.

[12] Patrick Finseth e Lasse Drevland. "Evaluating Machine Learning Methods for City Bike Demand Prediction in Oslo". Diss. de mestrado. NTNU, 2018.

[13] Andreas Kaltenbrunner et al. “Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system”. Em: *Pervasive and Mobile Computing* 6.4 (2010). Human Behavior in Ubiquitous Environments: Modeling of Human Mobility Patterns, pp. 455–466. ISSN: 1574-1192. DOI: <https://doi.org/10.1016/j.pmcj.2010.07.002>. URL: <http://www.sciencedirect.com/science/article/pii/S1574119210000568>.

[14] Junming Liu et al. “Functional Zone Based Hierarchical Demand Prediction For Bike System Expansion”. Em: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’17. Halifax, NS, Canada: ACM, 2017, pp. 957–966. ISBN: 978-1-4503-4887-4. DOI: <10.1145/3097983.3098180>. URL: <http://doi.acm.org/10.1145/3097983.3098180>.

[15] Junming Liu et al. “Rebalancing Bike Sharing Systems: A Multi-source Data Smart Optimization”. Em: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: ACM, 2016, pp. 1005–1014. ISBN: 978-1-4503-4232-2. DOI: <10.1145/2939672.2939776>. URL: <http://doi.acm.org/10.1145/2939672.2939776>.

[16] Hilary Mason. *A Taxonomy of Data Science*. 2010. URL: <http://www.dataists.com/2010/09/a-taxonomy-of-data-science/>. (accessado em 21.11.2019).

[17] Paul Philip Mitchell. “Predicting Bike-Sharing Traffic Flow using Machine Learning”. Diss. de mestrado. NTNU, 2018.

[18] Petrina Papazek et al. “A PILOT/VND/GRASP Hybrid for the Static Balancing of Public Bicycle Sharing Systems”. Em: fev. de 2013, pp. 372–379. DOI: <10.1007/978-3-642-53856-8_47>.

[19] Günther Raidl et al. “Balancing Bicycle Sharing Systems: Improving a VNS by Efficiently Determining Optimal Loading Operations”. Em: maio de 2013, pp. 130–143. DOI: <10.1007/978-3-642-38516-2_11>.

[20] Marian Rainer-Harbach et al. “Balancing Bicycle Sharing Systems: A Variable Neighborhood Search Approach”. Em: abr. de 2013, pp. 121–132. DOI: <10.1007/978-3-642-37198-1_11>.

[21] Tal Raviv e Ofer Kolka. “Optimal inventory management of a bike-sharing station”. Em: *IIE Transactions* 45 (out. de 2013), pp. 1077–1093. DOI: <10.1080/0740817X.2013.770186>.

[22] J. Schuijbroek, R.C. Hampshire e W.-J. van Hoeve. “Inventory rebalancing and vehicle routing in bike sharing systems”. Em: *European Journal of Operational Research* 257.3 (2017), pp. 992–1004. ISSN: 0377-2217. DOI: <https://doi.org/10.1016/j.ejor.2016.08.029>. URL: <http://www.sciencedirect.com/science/article/pii/S0377221716306658>.

[23] Susan A Shaheen, Stacey Guzman e Hua Zhang. “Bikesharing in Europe, the Americas, and Asia: past, present, and future”. Em: *Transportation Research Record* 2143.1 (2010), pp. 159–167.

[24] Susan A Shaheen et al. “China’s Hangzhou public bicycle: understanding early adoption and behavioral response to bikesharing”. Em: *Transportation Research Record* 2247.1 (2011), pp. 33–41.

[25] Susan Shaheen e Adam Cohen. “Shared Micromoblity Policy Toolkit: Docked and Dockless Bike and Scooter Sharing”. Em: (2019).

[26] Wen Wang. “Forecasting Bike Rental Demand Using New York Citi Bike Data”. Em: (2016).