Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Alexandre Ray da Silva, 01 de agosto de 2019

Predição de demanda de aluguel de bicicletas.

Histórico do assunto

O sistema de compartilhamento de meios de transporte é uma das alternativas de mobilidade urbana que vem ganhando cada vez mais espaço em grandes cidades como São Paulo. Mais conhecido como MaaS (Mobilidade-como-um-Serviço) representa uma mudança de propriedade pessoal de meios de transporte para soluções de mobilidade que são consumidos como um serviço.

No caso dos serviços de compartilhamento de bicicletas (bike sharing), elas são disponibilizadas para que qualquer pessoa possa utilizá-la por meio do desbloqueio a partir de um aplicativo de celular. Em geral, é permitido que o usuário alugue a bicicleta em um local e a deixe em outro lugar ou estação diferente da inicial conforme a necessidade de locomoção, proporcionando maior flexibilidade para os usuários. Além disso, a rede de sensores das bicicletas fornecem uma rica quantidade de dados que podem ser usados para estudo de mobilidade nas cidades.

Descrição do problema

O problema consiste em predizer a demanda pelo aluguel de bicicletas usando o histórico de demanda de bicicletas combinados com dados climáticos para o programa da Capital Bikeshare em Washington, D.C. A predição terá como referência o intervalo de uma hora. Dessa forma, queremos responder perguntas como:

1. Quantas bicicletas serão alugadas no dia 20/07/2019 às 15 horas?
2. E às 16 horas?

O conhecimento prévio da demanda pelo aluguel de bicicletas pode endereçar melhores estratégias de distribuição desse meio de transporte pela cidade, proporcionando melhores experiências para os usuários desses serviços.

Conjuntos de dados e entradas

O conjunto de dados do projeto pode ser coletado a partir da plataforma kaggle no endereço https://www.kaggle.com/c/bike-sharing-demand/data ou no endereço do UCI http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset. As variáveis disponibilizadas estão listadas abaixo:

Variável	Descrição
datetime	hourly date + timestamp
season	1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday	whether the day is considered a holiday
workingday	whether the day is neither a weekend nor holiday
weather	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp	temperature in Celsius
atemp	"feels like" temperature in Celsius
humidity	relative humidity
windspeed	wind speed
casual	number of non-registered user rentals initiated
registered	number of registered user rentals initiated
count	number of total rentals

A **cada hora**, são fornecidas informações do clima, se aquele dia é feriado ou não, temperatura, etc para um período de dois anos (2011 - 2012). A variável resposta é **count (Total de aluguéis).** Dessa forma, podemos criar um modelo para prever a quantidade total de aluguéis de bicicleta em uma determinada hora de acordo com os variáveis independentes fornecidas na tabela acima.

Fatores como temperatura, estação do ano e umidade parecem relevantes para a decisão de alugar ou não uma bicicleta. Provavelmente os usuários tendem a sair mais no verão e em temperaturas maiores do que no inverno, onde as pessoas preferem ficar em casa lendo um livro e tomando um café. Outro cenário pode ocorrer com feriados e dias não úteis. Provavelmente, nesses dias há maior demanda pelo alguel de bicicletas porque as pessoas têm mais tempo para passear.

A plataforma kaggle fornece três datasets para esse projeto:

1. sampleSubmission.csv (6493 observações)

2. Test.csv (6493 observações)
3. Train.csv (10886 obsevações)

O dataset 3 possui os dados e a variável resposta para os primeiros 19 dias de cada mês. E será usado para treinamento do modelo de aprendizado de máquina. Já os datasets 1 e 2 possuem as variáveis dependentes e independentes respectivamente para o período do 20º dia de cada mês até o final do mesmo.

Descrição da solução

Nesse projeto, usaremos técnicas de Aprendizado de Máquina supervisionado. As nossas variáveis dependentes serão as variáveis como temperatura e clima (fatores que influenciam o aluguel de bicicletas). E a variável resposta será o número total de bicicletas alugadas para cada hora. Usaremos modelos baseados em regressão já que queremos predizer um valor contínuo (número de bicicletas alugadas). Os modelos que serão usados e a métrica de avaliação, bem como o benchmark serão detalhados nas próximas seções.

Modelo de referência (benchmark)

O modelo de referência escolhido será o Kernel com maior número de votos para esse projeto na plataforma do Kaggle. Portanto, é possível utilizar como benchmark o resultado dos seguintes modelos de aprendizado de máquina para este kernel: *LinearRegression, Ridge, Lasso, RandomForestRegressor e GradientBoostingRegressor*. A métrica de avaliação chave será a RMSLE e será explicada mais detalhadamente na próxima seção.
Link: https://www.kaggle.com/viveksrinivasan/eda-ensemble-model-top-10-percentile

Métricas de avaliação

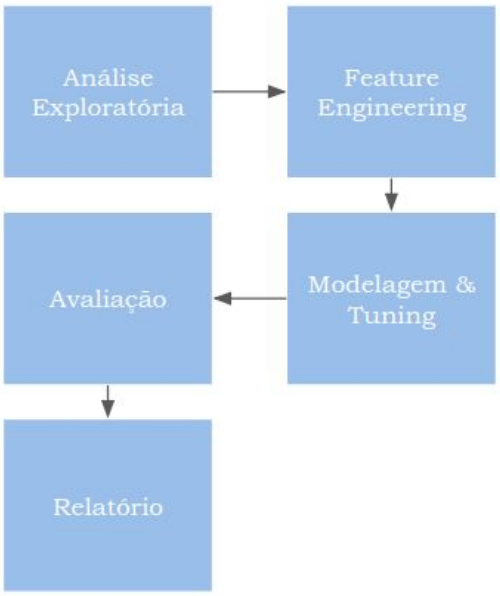
A métrica de avaliação a ser utilizada nesse projeto é a Root Mean Squared Logarithmic Error (RMSLE). A métrica RMSLE geralmente é usada quando não queremos penalizar muito as diferença entre o valor real e o valor predito no caso em que o valor real e predito são muito grandes.

Esse modo de lidar com valores muito grandes é interessante para o problema aqui proposto. Por exemplo, quando a demanda de bicicletas atinge um valor considerado muito alto, o tamanho do erro passa a não importar mais. Em outras palavras, a demanda de 10k bicicletas em uma determinada região já não é tão diferente de 100K bicicletas uma vez que a situação já pode ser considerada crítica (altíssima demanda) na primeira situação. Matematicamente a métrica RMSLE é descrita como:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

Design do projeto

O fluxo de trabalho adotado está descrito no diagrama abaixo:



Análise Exploratória: Nessa etapa, serão analisadas detalhadamente cada uma das variáveis fornecidas a fim de compreender melhor o problema e identificar quais variáveis serão melhores preditoras. Além disso, análise de missing values, outliers e a distribuição da variável resposta são indispensáveis nessa fase.

Feature Engineering: A partir das análises realizadas na etapa anterior, possivelmente chegarei em algumas transformações das variáveis originais que farão maior sentido para a solução do problema. Esta etapa consiste em construir essas variáveis transformadas para que sejam usadas no modelo.

Modelagem & Tuning: Nessa etapa, serão aplicados os algoritmos de aprendizado de máquina supervisionado (*LinearRegression, Ridge, Lasso, RandomForestRegressor e GradientBoostingRegressor*). Além disso, serão testadas técnicas de tuning do modelo.

Avaliação: Uma vez que os modelos foram criados, essa etapa consiste em avaliar a performance de cada um dos algoritmos, bem como compará-la com a performance do modelo de benchmark proposto.

Relatório: Será feito um relatório detalhado sobre todo o processo de desenvolvimento do modelo e como ele pode ser usado para resolver o problema de negócio.