# \*\*Situação Problema

Uma empresa de táxi da cidade de Nova York deseja desenvolver um aplicativo para prever a duração das viagens que serão feitas. Para isso, foi disponibilizado um conjunto de dados de viagens de New York City feitos por taxi e limousines, que inclui tempo de coleta, geocoordenadas, número de passageiros e entre outras variáveis.

Com base nisso, este projeto desenvolverá um algoritmo de regressão de aprendizado de máquina capaz de prever a duração de uma viagem com base nas variáveis fornecidas pelo usuário do aplicativo (passageiro) e pelo motorista.

Neste notebook serão realizadas as seguintes etapas:

- Introdução: Visa esclarecer o objetivo do projeto e as etapas a serem seguidas.

- Limpeza e pré-processamento dos dados: Será feita uma pré-visualização dos dados, a fim de identificar e tratar dados inconsistentes e ausentes, além de criar recursos extras com base nos recursos iniciais.

- Análise Exploratória de Dados (EDA): Serão realizadas visualizações para encontrar insights relevantes para o desenvolvimento do algoritmo.

- Correlação: A correlação entre os dados será analisada a fim de evitar possíveis problemas de multicolinearidade.

- Implementação do modelo: É a etapa onde será criado e avaliado o modelo de aprendizado de máquina, onde serão ajustados os hiperparâmetros do modelo para melhorar o resultado do modelo, e será realizada a análise de importância de features do modelo final.

- Conclusão: Etapa onde serão feitos comentários a respeito do projeto realizado e também alguns levantamentos de melhorias futuras para o projeto.

# \*\*Limpeza e pré-processamento dos dados\*\*

<img src="/New york city - taxi trip duration/project\_prints/desc\_var.png ">

É possível observar no dataset 6 variáveis numéricas, 3 variáveis categóricas e 2 variáveis ‘DateTime’.

<img src="/New york city - taxi trip duration/project\_prints/data\_null.png ">

Observando, vemos que inicialmente não há dados faltantes no conjunto de dados de treinamento.

Observe que as variáveis 'pickup\_datetime' e 'dropoff\_datetime' estão no formato de objeto em vez de datetime, exigindo uma alteração de formato.

Para um melhor entendimento da distância percorrida entre o embarque e o desembarque, será criada uma nova variável que representará a \*\*distância\*\* aproximada entre essas coordenadas geográficas em metros. também será calculado a \*\*direção do vetor\*\* formado pelos pontos de embarque e desembarque.

Para uma melhor representação do 'pickup\_datetime' será extraído o \*\*horário\*\*, \*\*dia da semana\*\* e \*\*mês\*\* em que a viagem foi solicitada.

# \*\*Análise Exploratória dos Dados (EDA)\*\*

<img src="/New york city - taxi trip duration/project\_prints/img\_01.png ">

Analisando a tabela temos:

- Em média as viagens têm de 1 a 2 passageiros.

- Observe que o menor número de passageiros é 0, portanto, existem alguns registros que indicam que a viagem foi para transporte de objetos ao invés de pessoas, ou indicam que existem registros que foram preenchidos incorretamente.

- O maior número de passageiros transportados ao mesmo tempo foi de 9 pessoas.

- Em média, uma viagem dura 959,49 segundos (aproximadamente 16 minutos).

- Observando o menor valor da duração da viagem, encontra-se o valor de 1 segundo, sendo um valor inconsistente para uma viagem.

- O tempo mais longo de uma viagem foi de 3,526\*10^6 segundos, aproximadamente 41 dias, representando uma viagem para fora da cidade ou dados inconsistentes.

- Em média a viagem tem uma distância de 3,94 km.

- Nota-se também que existem valores de distância percorrida iguais a 0, indicando que não ocorreu uma viagem ou que o destino da viagem foi o mesmo da partida.

- A viagem mais longa teve uma distância de 1243,4 km, que pode representar uma viagem para fora da cidade.

- Observe que o maior valor de 'month\_pickup' é 6 e o menor é 1, indicando que os dados são dos meses entre janeiro e junho do ano analisado.

Como ponto de partida será analisado as variáveis 'trip\_duration' (target) e 'distance' buscando identificar o motivo de alguns valores inconsistentes observados na tabela acima.

<img src="/New york city - taxi trip duration/project\_prints/img\_02.png ">

No gráfico de densidade da variável target, observa-se um formato assimétrico à direita, visivelmente com um skewness positivo elevado.

Como a concentração destes valores tende a ser pequena, a representação na forma logarítmica pode resultar em uma melhor visualização destes dados.

<img src="/New york city - taxi trip duration/project\_prints/img\_03.png ">

Através da representação logarítmica, o gráfico passa a ter um formato mais parecido com uma distribuição normal, tendo a maior concentração de dados entre os valores de 4 e 9 (representando valores entre 55 e 8100 segundos).

Existe também um pequeno pico de densidade entre os valores de 11 e 12 na escala logarítmica (valores entre 59900 e 162800 segundos), que são valores intuitivamente muito altos para uma viagem.

Existem pequenas ondulações para valores menores que 3 na escala logarítmica (tempos de viagem menores que 20 segundos), que são valores inconsistentes de tempo de viagem.

<img src="/New york city - taxi trip duration/project\_prints/img\_04.png ">

analisando o gráfico de densidade da variável 'distance' nota-se que ela possui o mesmo problema que 'trip\_duration'. Com base nisso, será feita uma visualização em formato logarítmico.

<img src="/New york city - taxi trip duration/project\_prints/img\_05.png ">

No gráfico de densidade, nota-se certa semelhança com uma distribuição normal, sendo um pouco diferente para valores de distância logarítmica entre 9 e 10, onde há um pico de concentrações na mesma distância.

Observe também uma concentração de dados de distância próxima ou igual a 0 (que devido a conversão em log(distance) +1 é aproximadamente 0 metros), sendo valores muito inconsistentes para uma viagem.

Para uma melhor visualização dos dados em análises subsequentes as variáveis 'distance' e 'trip\_duration' serão representadas em formato logarítmico.

<img src="/New york city - taxi trip duration/project\_prints/img\_06.png ">

Pode-se observar no gráfico que mais de 90% dos dados são de viagens com até 3 passageiros, sendo que quase 71% das viagens são feitas com um único passageiro.

Observa-se a presença de dados com número de passageiros igual a 0, mas com valor próximo de 0%.

<img src="/New york city - taxi trip duration/project\_prints/img\_07.png ">

Analisando o gráfico de coordendas de pickup:

- A maioria dos pontos está concentrada em uma região que pode ser considerada a cidade principal onde esta franquia de taxi atua.

- Existem alguns pontos muito distantes desta cidade, que podem ser consideradas viagens atípicas.

Visando agregar mais informação na elaboração do modelo será realizado uma clusterização para os ponstos de pickup e de dropoff, visando separar regiões mais comuns das regiões mais atípicas das viagens.

Para isto será utilizado o modelo de aprendizagem 'Kmeans', devido a sua capacidade de realizar agrupamentos em uma larga quantidade de dados. Para isto, será utilizado como critério de decisão de número de clusters o método da inércia.

<img src="/New york city - taxi trip duration/project\_prints/img\_08.png ">

Observando a curva de cotovelo, o valor aproximado de clusters para os pontos de embarque é 6.

<img src="/New york city - taxi trip duration/project\_prints/img\_09.png ">

Visualmente os grupos parecem estar bem separados, com excessão do grupo 2 que aparenta ter poucos dados.

<img src="/New york city - taxi trip duration/project\_prints/img\_10.png ">

Observando o gráfico da curva de cotovelo, nota-se que o valor ideal de clusters é 8 para o ponto de desembarque.

<img src="/New york city - taxi trip duration/project\_prints/img\_11.png ">

Pelo gráfico nota-se que a maioria dos grupos está bem separados, com excessão do grupo 1 que aparenta estar com poucos dados.

<img src="/New york city - taxi trip duration/project\_prints/img\_12.png ">

Observe que para o recurso 'store\_and\_fwd\_flag' a maioria das viagens os veículos tinham conexão com a internet (classe N).

A partir do gráfico de 'vendor\_id' pode-se ver que esse recurso está bem balanceado, mas possui uma quantidade maior de id 2.

<img src="/New york city - taxi trip duration/project\_prints/img\_13.png ">

Pelos gráficos não é possível identificar diferença significativa na distribuição dos dados de 'vendor\_id' em relação a 'trip\_duration\_log' e 'distance\_log'.

Observe que o 'vendor\_id 1' lida com viagens para até 4 pessoas, e raramente atende viagens com uma quantidade maior, enquanto o fornecedor 2 lida com viagens para qualquer número de pessoas.

<img src="/New york city - taxi trip duration/project\_prints/img\_14.png ">

Pelo gráfico do 'month\_pickup' pode-se observar que todos os meses apresentam valores próximos à quantidade de dados, sendo o menor valor em janeiro e o maior em março.

Pelo gráfico de 'hour\_pickup' pode-se observar que tem uma menor demanda de viagens entre 00h00 e 05h00 e uma demanda maior entre 17h00 e 22h00.

A partir do gráfico de 'day\_week\_pickup' é possível observar que os dias da semana não possuem diferença relevante na quantidade de dados. Segunda e domingo são os dias com menos viagens e sábado e sexta são os dias com mais quantidade de viagens.

<img src="/New york city - taxi trip duration/project\_prints/img\_15.png ">

Pelo gráfico é possível observar que:

- As viagens realizadas entre 13h e 17h têm duração média superior a 1000 segundos, tendo o pico ás 15 horas.

- As viagens feitas entre 4h e 8h têm a menor duração média de viagem.

- Entre o período das 6h00 às 13h00 há um aumento gradativo na duração média da viagem.

- Entre o período das 17:00 às 19:00 há uma diminuição gradual na duração média da viagem.

Com base nessas características, é possível inferir que esse recurso possui características importantes que podem ser relevantes na construção do modelo de previsão da 'trip\_duration'.

<img src="/New york city - taxi trip duration/project\_prints/img\_16.png ">

É possível identificar pelo gráfico:

- Existe uma relação crescente entre a média de duração da viagem e os dias de segunda a quinta-feira.

- Existe uma relação decrescente entre a média de duração da viagem e os dias de quinta a domingo.

- Os dias com os menores valores médios de duração da viagem são o domingo e a segunda-feira.

<img src="/New york city - taxi trip duration/project\_prints/img\_17.png ">

Existe uma relação linear crescente entre os meses e o valor médio da duração das viagens.

Como o ciclo de um ano de análise dos dados não foi concluído, pode ser que essa informação não seja tão relevante, pois pode ser um efeito cíclico e não um efeito crescente ao longo dos meses.

<img src="/New york city - taxi trip duration/project\_prints/img\_18.png ">

Analisando o gráfico nota-se:

- Existe uma região com uma maior concentração de dados, onde existe uma certa linearidade entre as duas variáveis.

- A faixa entre 0 e 1 para a variável 'distance' na escala logarítmica apresenta uma variedade de valores de 'trip duration' muito ampla, porém os valores de 'distance' aparentam estar 'padronizados', não possuindo um valor intermediário aleatório entre eles. Isto pode indicar algumas situações, como: Tour pela cidade (com retorno ao ponto de partida ou próximo), travessias de pontes ou rios ou engarrafamentos.

- No valor próximo de 11 em 'trip duration' na escala logarítmica (valor de aproximadamente 16 horas) ocorre uma concentração de diferentes valores de 'distance'. Isto pode ter sido gerado por algum problema climático, obstrução de vias para eventos, etc...

- A faixa entre 0 e 2 para a variável 'trip duration' na escala logarítmica apresenta uma variedade de valores de 'distance' ampla, porém os valores de 'trip duration' aparentam estar 'padronizados', não possuindo um valor intermediário aleatório entre eles. Isto pode indicar que estas viagens foram canceladas ao invés de terem sido realizadas.

#\*\*Correlação\*\*

Para evitar problemas de multicolinearidade será avaliado a correlação entre as variáveis.

Como não existe uma literatura muito clara para valores de correlação 'aceitáveis' que evitem a multicolinearidade, valores maiores que 0,7 de correlação serão definidos como valores de corte a serem removidos.

<img src="/New york city - taxi trip duration/project\_prints/img\_19.png ">

Somente a variável 'dropoff\_longitude' possui correção maior que 0,7 com a variável 'pickup\_longitude', portanto 'dropoff\_longitude' será removido antes da elaboração do modelo.

# \*\*Implementação do modelo\*\*

Como modelo de regressão será utilizado o \*\*LightGBM\*\*, sendo um modelo de aprendizado de máquina que possui um ótimo desempenho para resolução de problemas de regressão. Como métrica principal será utilizado a \*\*Raiz do Erro Médio Quadrado Logarítmico (RMSLE)\*\*, visando penalizar menos erros elevados em valores altos de tempo previstos.

Antes da aplicação do modelo de regressão, será aplicada a normalização dos dados quantitativos, visando aumentar o desempenho do modelo e evitar que o algoritmo desenvolva viés para variáveis de ordem de grandeza superior.

Para os dados categóricos será aplicado o método 'OneHotencoder' para a transformação destas variáveis.

Durante o aprendizado do modelo, o método de validação cruzada será aplicado, a fim de avaliar a generalização do modelo.

Com o objetivo de obter o melhor desempenho do modelo será aplicado o hiper tunning dos hiperparâmetros do modelo LightGBM

<img src="/New york city - taxi trip duration/project\_prints/img\_20.png ">

Com o processo de otimização, observa-se uma melhora pequena no valor do RMSLE ao decorrer da otimização, resultando um RMSLE de 0.3112 para o melhor modelo.

<img src="/New york city - taxi trip duration/project\_prints/img\_21.png ">

Dos gráficos de variação dos parâmetros, nenhuma variável tendeu aos limites inicialmente estabelecidos à medida que ocorreu a otimização, excluindo a variável 'colsample\_bytree', cujo valor máximo possível é 1. Isso é uma indicação de que a otimização foi realizada corretamente e nenhum parâmetro precisa ser expandido em uma nova otimização para melhorar ainda mais o desempenho do modelo.

<img src="/New york city - taxi trip duration/project\_prints/img\_22.png ">

Observa-se que a variável 'learning\_rate' foi a variável que mais agregou valor para melhorar o desempenho do modelo. As outras variáveis como 'subsample', 'n\_estimators' e 'num\_leaves' tiveram um impacto menor e poderiam ser removidas para reduzir o tempo de otimização.

<img src="/New york city - taxi trip duration/project\_prints/img\_23.png ">

Neste caso, todos os dados para treinamento do modelo são utilizados para aumentar o desempenho do algoritmo gerado, aumentando assim os resultados com os dados finais do teste.

Observe que houve uma melhora de aproximadamente 0,06 no valor de RMSLE em relação aos resultados com validação cruzada.

# \*\*Coclusão\*\*

O resultado obtido com o algoritmo ficou próximo do resultado esperado e espera-se que quando aplicado aos dados de teste sejam obtidos resultados semelhantes.

A fim de melhorar o projeto em trabalhos futuros, as seguintes atividades podem ser realizadas:

- Remoção de Features que adicionam menos informações ao modelo, verificando o aumento de performance.

- Realizar uma melhor limpeza nos dados das coordenadas de embarque e desembarque, para agregar maior poder preditivo.

- Aplicação de outros modelos de regressão para comparação de desempenho.