Previsão de pontos finais de trajetórias de táxi em Porto - Portugal

CAIO CASAGRANDE

Introdução

Sobre o Negócio:

- Com as novas tecnologias atuais, a indústria de táxis precisou se **reinventar** para não ficar para trás em relação aos seus novos concorrentes
- Um dos **desafios é a adaptação** ao novo sistema eletrônico de despacho em tempo real, instalados nos veículos

Introdução

Sobre o Problema:

- O sistema conta com um problema: **a falta de informação sobre o destino final das corridas**, pois os motoristas não indicam o destino;
- Os despachantes precisam identificar corretamente qual táxi enviar para uma localização de coleta, o que se torna difícil quando não se sabe o destino final dos táxis em serviço;
- Em razão desse problema, a proposta é **desenvolver um modelo preditivo** que seja capaz de inferir o destino final de corridas de táxi com base em suas localizações de coleta.

Objetivos

- Entender o problema de negócio
- Análise de Dados
- Modelo de Machine Learning
- PowerPoint descrevendo o problema e conclusões
- Enviar notebook (.ipynb) e apresentação (.pdf)

Planejamento

- Saídas: Modelo, Arquivo .ipynb, Apresentação;
- Entrada: Dados
- Passo a passo:
  - Data Overview
  - Pré-processamento
  - Análise Exploratória dos Dados
  - Feature Engineering
  - Modelagem Machine Learning
  - Avaliação
  - Performances
    - Modelo
    - Negócio

Data Overview

- Dataset com 1,710,670 linhas e 9 colunas
- Duas variáveis com dados faltantes: ORIGIN_CALL e ORIGIN_STAND

Variable	Description	Type
TRIP_ID	Identificação de cada viagem.	int
CALL_TYPE	Identifica a maneira que o serviço aconteceu (A, B ou C)	object
ORIGIN_CALL	Identificação de número de telefone que pediu táxi	float
ORIGIN_STAND	Ponto de táxi em que o pedido foi realizado	float
TAXI_ID	Identificação do Táxi	int
TIMESTAMP	Timestamp indicando quando ocorreu a corrida	int
DAY_TYPE	Indica o tipo de dia (A, B ou C)	object
MISSING_DATA	Indica se há falta de dados	bool
POLYLINE	Sequência de coordenadas geográficas do trajeto	object

Pré-processamento

- Transformação dos nomes das colunas para *snake*_case
- Transformação da variável "*timestamp*" de segundos para um estilo de data
- Novas variáveis de tempo (**hora**, **dia**, mês, ano, dia da semana, semana do ano)
- Criação de colunas com **nomes** para dias da semana e meses
- **Filtragem** de dados:

  Ano == 2013

  Meses de Julho a Novembro
- Selecionando apenas linhas em que não havia dados faltantes

  'missing_data' == False

Pré-processamento

- Variável ‘Polyline’ de *string* para lista de elementos (**coordenadas**)
- Criação da variável ‘distance’ utilizando **Haversine**
- Excluindo distâncias muito grandes (> 0.975) e muito pequenas (<0.025)

Análise Exploratória dos Dados

3.1. Quais os tipos de chamadas mais frequentes?

3.2. Quais são os telefones que mais solicitaram corridas de táxi?

3.3. Quais os pontos de táxi com origem mais frequentes?

3.4. Quais os taxistas com maior número de viagens?

3.5. Quais taxistas percorreram mais distância em suas viagens?

3.6. Qual a distribuição das viagens por dia?

3.7. Existe algum comportamento sazonal dentro dos meses?

3.8. Qual o dia da semana com mais corridas? Qual dia se percorre as maiores distâncias?

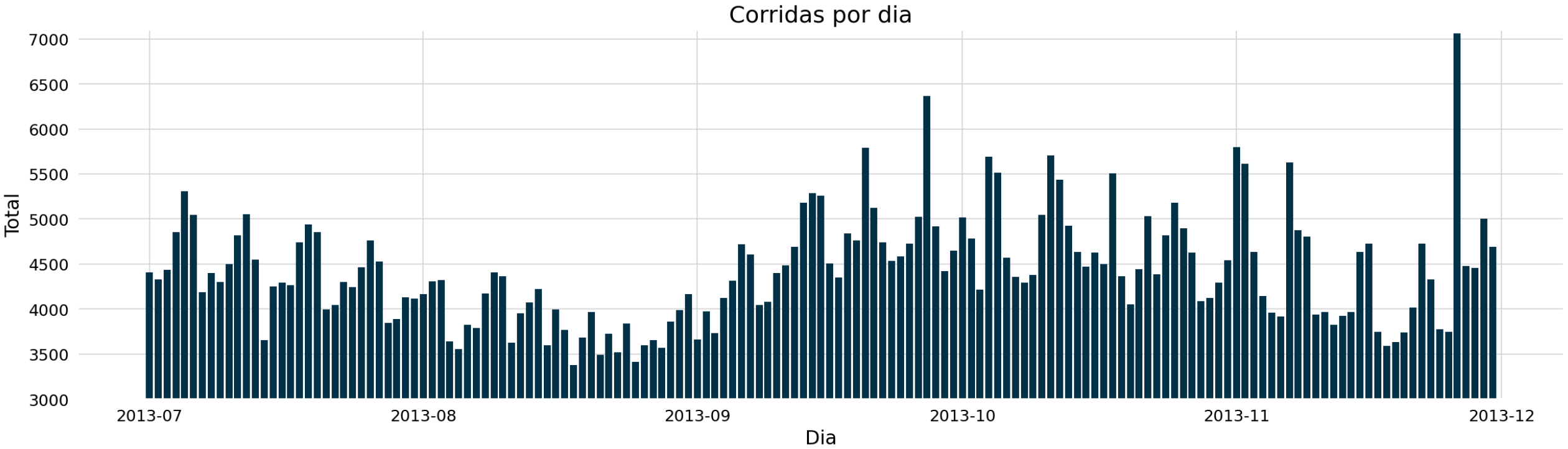
3.9. O comportamento das distâncias percorrida muda de acordo com o tipo de chamada?

3.10. Qual o comportamento das distâncias percorridas nas semanas do ano?

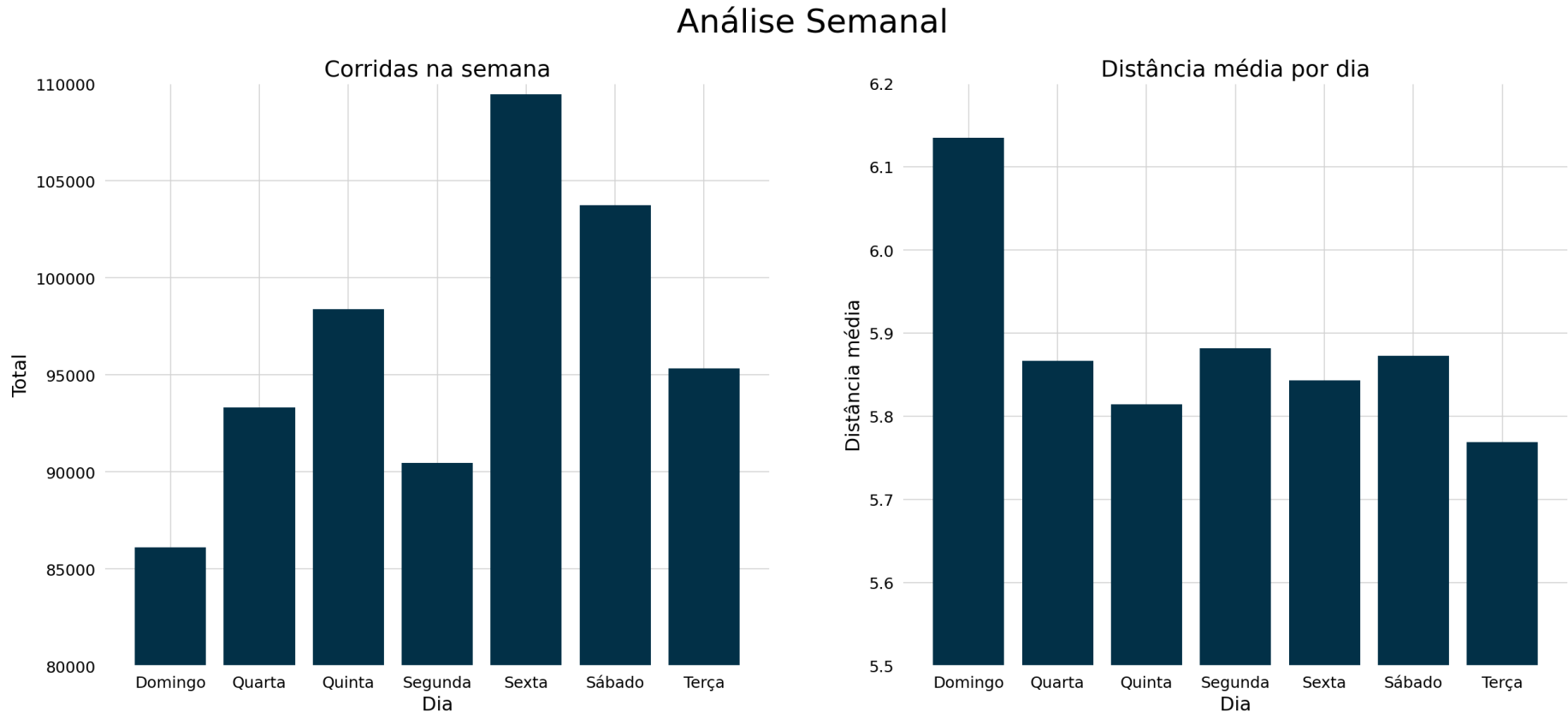
3.11. Qual o comportamento das distâncias percorridas por mês?

3.12. Qual o comportamento das distâncias percorridas por hora do dia?

Análise Exploratória dos Dados

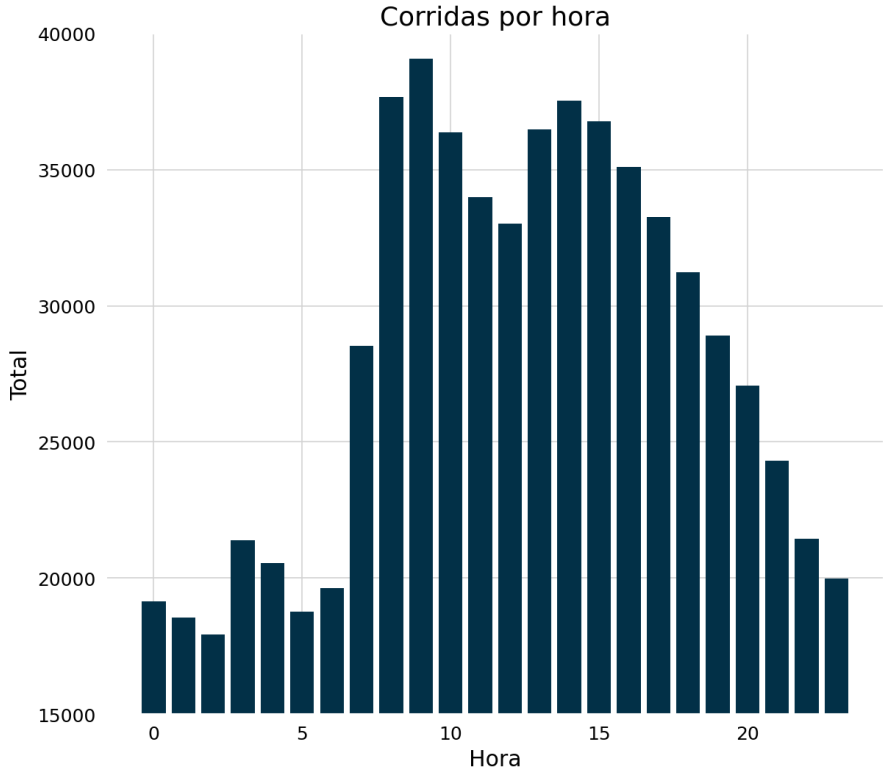


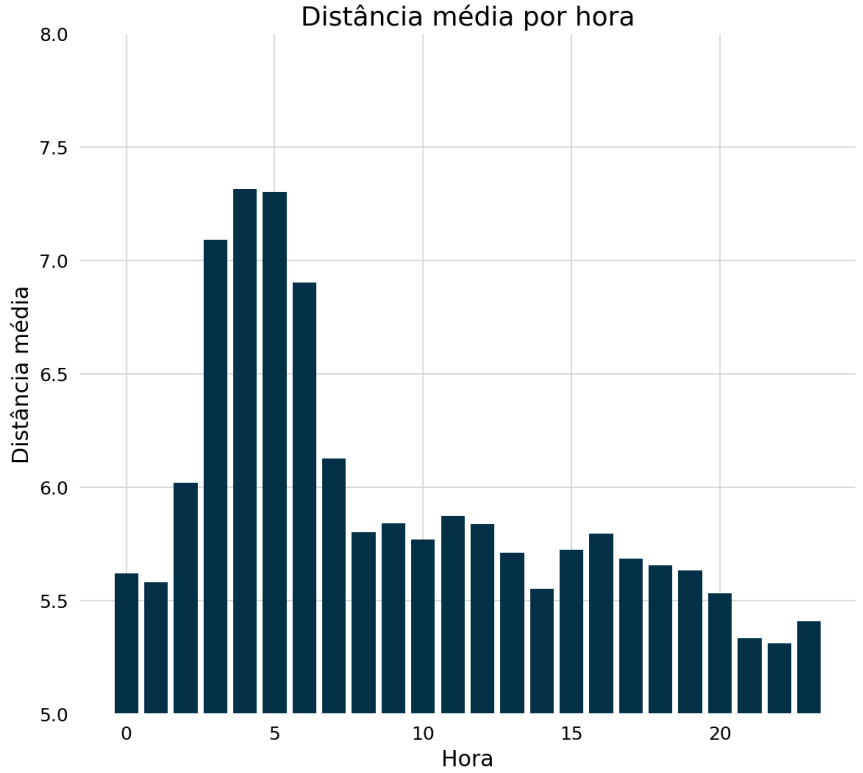
Análise Exploratória dos Dados



Análise Exploratória dos Dados

Análise diária





Feature Engineering

- Criação de novas variáveis: posições iniciais e finais de latitude e longitude
- Exclusão de colunas desnecessárias para o modelo:

  - `trip_id` : um identificador único para cada corrida não influencia no modelo;
  - `day_type` : todos os valores da coluna são iguais a "A";
  - `taxi_id` : o identificador único de um táxi não influencia no ponto final de chegada da corrida;
  - `timestamp` : datetime feature
  - `polyline` : o trajeto é um resultado final de cada corrida, de maneira que os trajetos passados não são relevantes para prever o ponto final de uma corrida futura;
  - `missing_data` : todos os valores da coluna são iguais a "False";
  - `name_dayofweek` : object feature;
  - `name_month` : object feature;
  - `weekofyear` : não é relevante para prever corridas em semanas que ainda não aconteceram;
  - `day_month_year` : datetime feature.

Feature Engineering

- One-Hot Encoding para coluna “call_type”
- Substituindo resultados faltantes em “origin_call” e “origin_stand” por zeros
- Sem riscos de *data leakage*
- Train-test split com amostra de tamanho 250 mil

Machine Learning

- LightGBM: **lightgbm.LGBMRegressor**
- Considerando que se trata de um grande dataset
- O modelo é reconhecido pela sua boa performance, rapidez com grandes datasets e eficiência de memória computacional.

Machine Learning

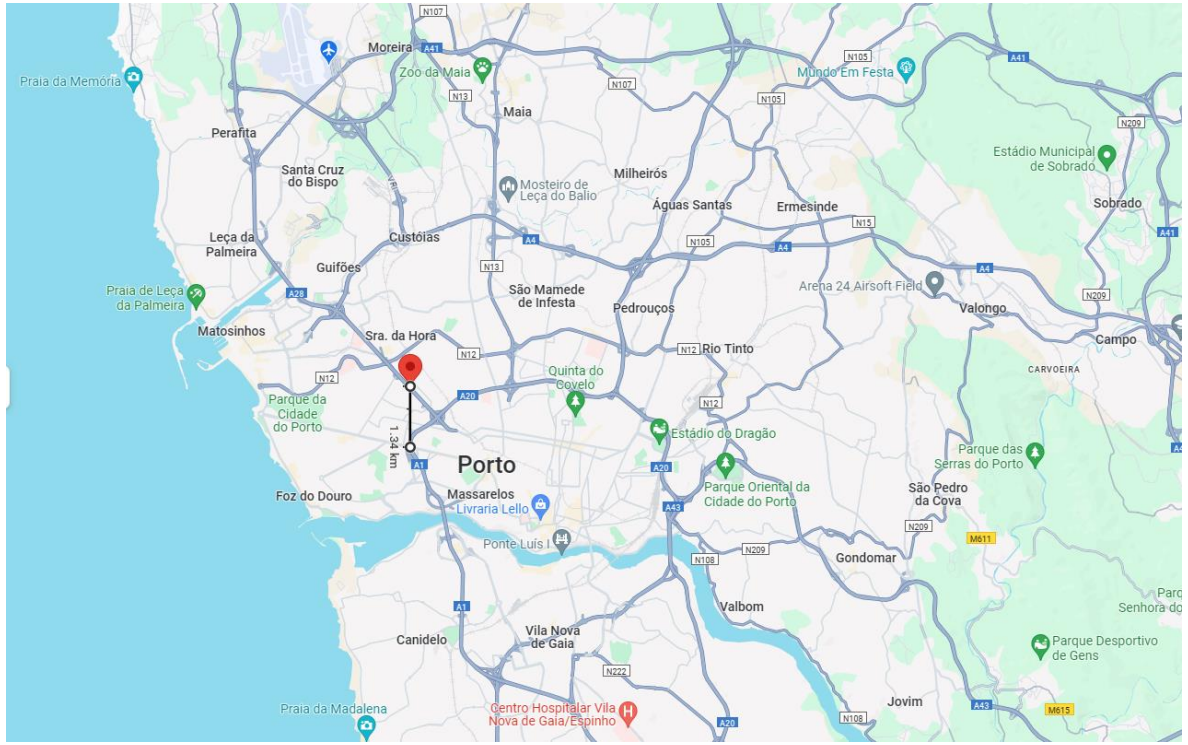
- Primeiro modelagem sem realizar alterações em parâmetros
  - Modelo
  - Predição
  - Avaliação
- Segunda modelagem com GridSearchCV

Primeiro Modelo
Mean Squared Error (Latitude): **0.00036**
Mean Absolute Error (Latitude): **0.01227**
Mean Squared Error (Longitude): **0.00067**
Mean Absolute Error (Longitude): **0.01880**

Modelo GridSearchCV
Mean Squared Error (Latitude): **0.00035**
Mean Absolute Error (Latitude): **0.01192**
Mean Squared Error (Longitude): **0.00065**
Mean Absolute Error (Longitude): **0.01828**

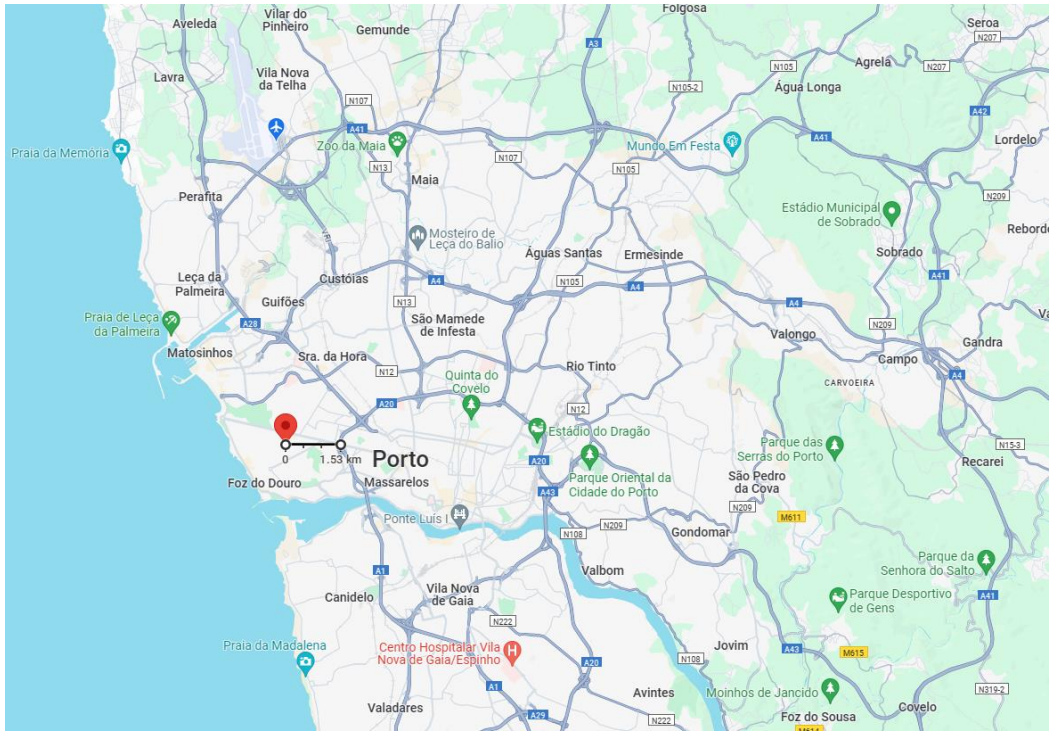
Machine Learning

Mean Absolute Error (Latitude): **0.01192 ~ 1.34 km**



Machine Learning

Mean Absolute Error (Longitude): **0.01828 ~ 1.53 km**



Machine Learning

- Modelo salvo em *pickle*
- Rodando no dataset de teste
- **MAE Latitude** 1.34 km – **0.83 km**
- **MAE Longitude** 1.53 km – **1 km**

Dataset de Teste
Mean Squared Error (Latitude): **0.00020**
Mean Absolute Error (Latitude): **0.00795**
Mean Squared Error (Longitude): **0.00030**
Mean Absolute Error (Longitude): **0.01238**

Performance

- A base utilizada para a modelagem foi apenas uma fração do total em razão de limites computacionais;
- Acredita-se que utilizando a base inteira, o modelo **LightGBM** possa prever com mais exatidão os pontos de destino;
- Ademais, incrementar o **GridSearchCV** com mais parâmetros também pode resultar em modelos mais precisos;
- Ainda assim, **o modelo resultou em bons resultados para a base de treino e resultados ainda melhores para o dataset de teste (erros menores).**

Performance

Realizar a previsão do destino final de corridas de táxi com base em pontos iniciais pode trazer diversos benefícios financeiros à empresa, como:

1. Minimização de quilometragem vazia
2. Maior satisfação do cliente
3. Otimização de recursos
4. Mais viagens em menos tempo
5. Planejamento operacional
6. Vantagem competitiva

Previsão de pontos finais de trajetórias de táxi em Porto - Portugal

CAIO CASAGRANDE