Apresentação Taxi_Drive

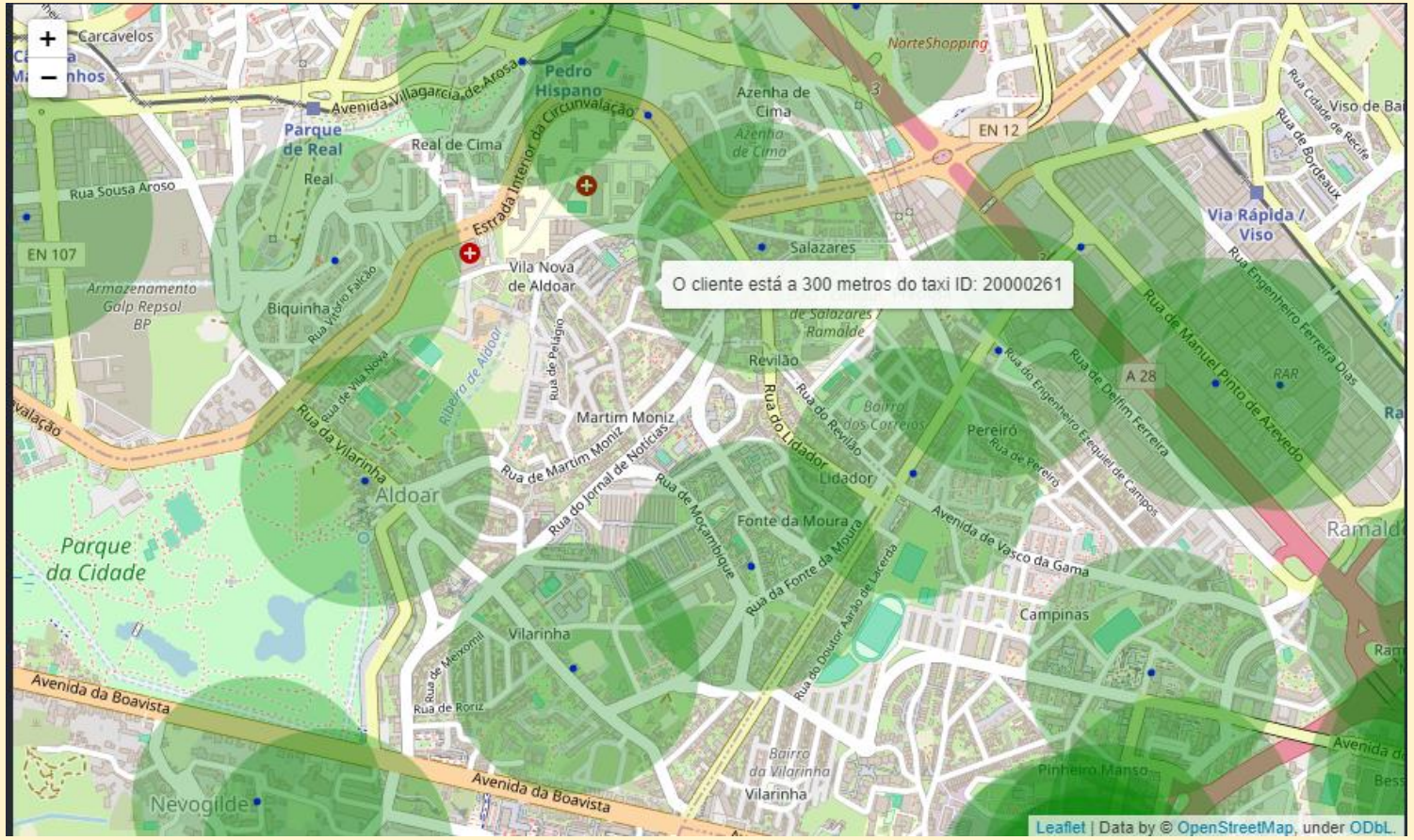
Cientista de Dados Gabriel Nobre

Problema de negócio

- Necessidade de modernização na área de locomoção urbana via táxi.
- Construção de um modelo capaz de prever o destino final das corridas com base na sua trajetória inicial.
- Se um despachante soubesse aproximadamente onde seus motoristas de táxi terminariam suas viagens atuais, ele seria capaz de identificar qual táxi atribuir a cada solicitação de coleta.
- Neste desafio, pedimos-lhe que construa uma estrutura preditiva que seja capaz de inferir o destino final das corridas de táxi no Porto, Portugal, com base nas suas trajetórias parciais (iniciais).

Solução Proposta

- A solução será um modelo em produção com a possibilidade do operador informar lat e long inicial do taxista e obter o possível lat e long final dentro de um mapa com uma área delimitada para encontrar o táxi mais próximo do cliente, conforme o mapa ao lado.



Trabalho com os dados

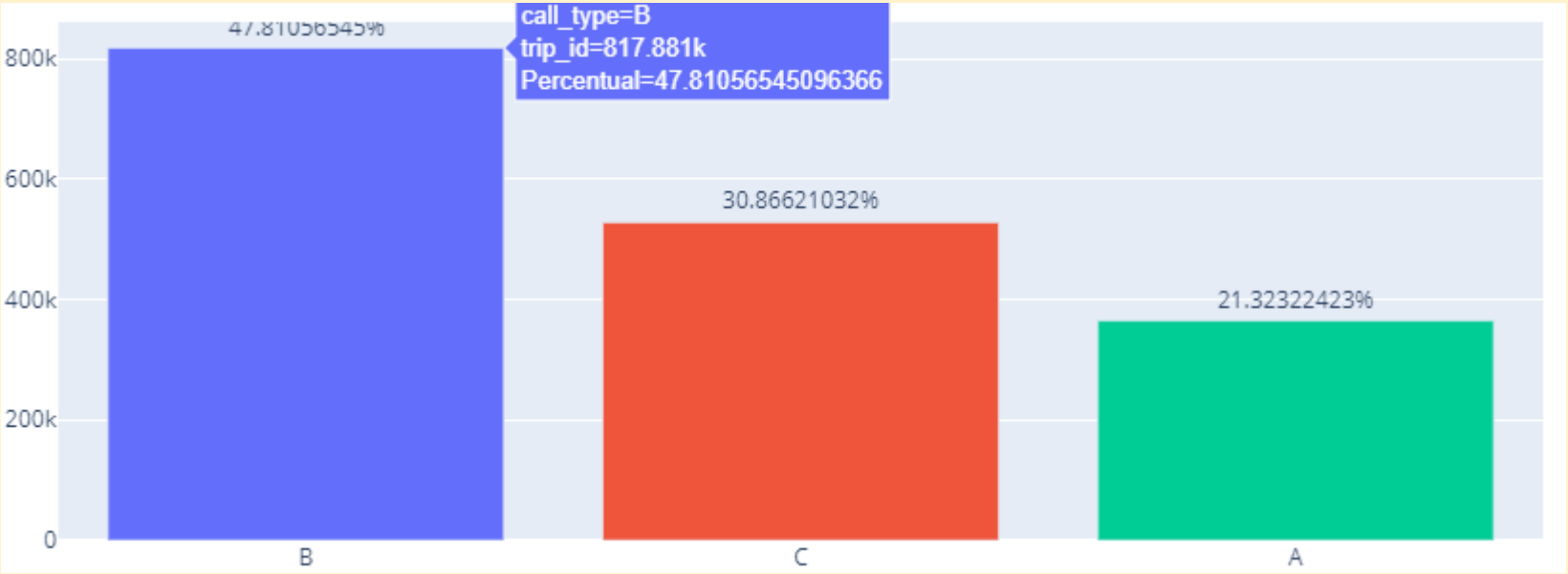
- A tabela base possuía 9 colunas, sendo: 3 de texto, 1 boleana e 5 numéricas.
- Tratei os NA’s da colunas origin_call e origin_stand, conforme modelo de negócio.
- As colunas que deram mais trabalho de limpar e extrair valores foram timstemp e polyline. Foram as features derivadas dessas colunas que mais contribuíram dom o modelo ao longo do tempo. Daqui consegui extrair lat e long inicial e final, no qual são essenciais para o modelo.
- Foi calculado a métrica de Haversine e colocada junto a tabela.

EDA e Hipóteses de negócio

- Foram levantas 7 hipóteses ao total, irei trazer 4 de maior impacto:
- 1. Há mais viagens com o fluxo de dados incompletos do que completos. Hipótese falsa.
- Só há 10 viagens que estão com o fluxo incompleto. Podendo ser criada novas features:
  - 1. "tempo_da_viagem" = Calcular o tempo total da viagem em minutos. Cada ping equivale a 15 segundos.
  - 2. "dt_hora_viagem_final" = Calcular a data e hora final da viagem.

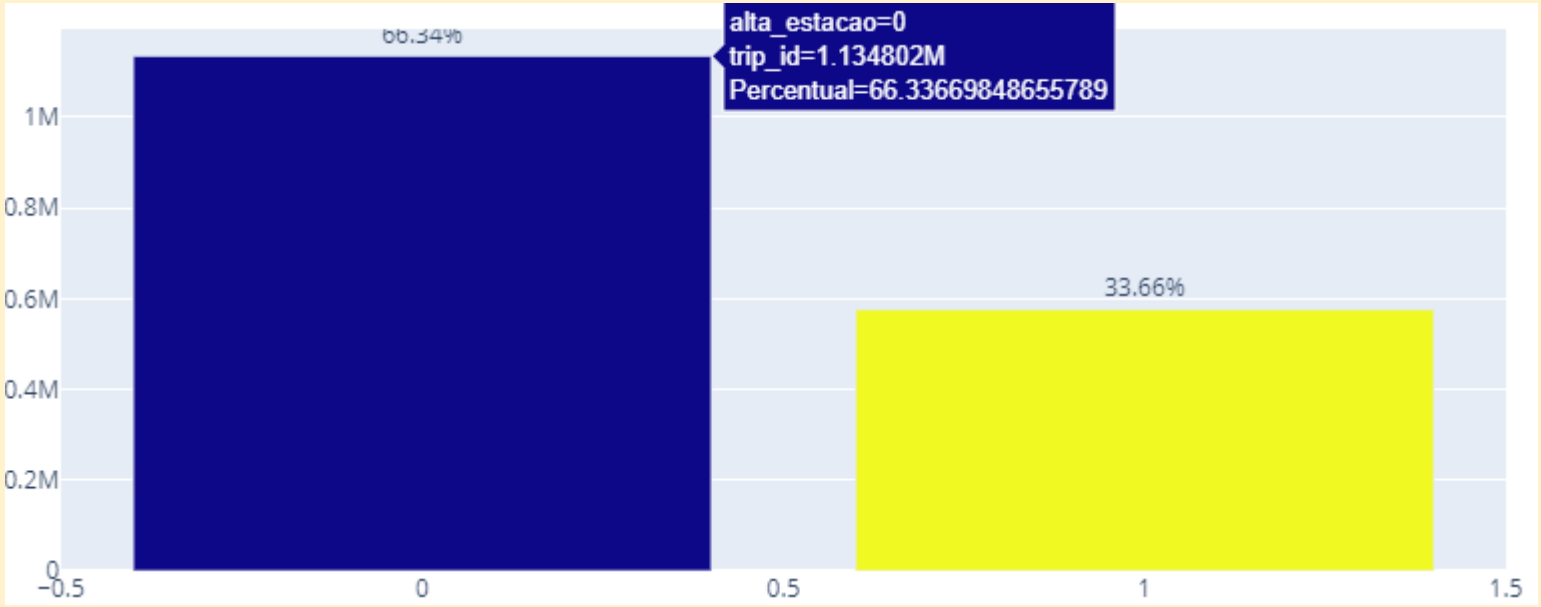
EDA e Hipóteses de negócio

- 4. As viagens despachadas da central não atingem 30% das viagens totais. Hipótese Verdadeira.
- 5. Pelo menos 50% das viagens são pegues em pontos específicos. Hipótese Verdadeira.
  - É interessante um estudo de viabilidade da construção de um aplicativo para que o cliente possa pedir seu taxi sem precisar nescessariamente passar pela central, mas passando por dentro da estrutura da empresa.



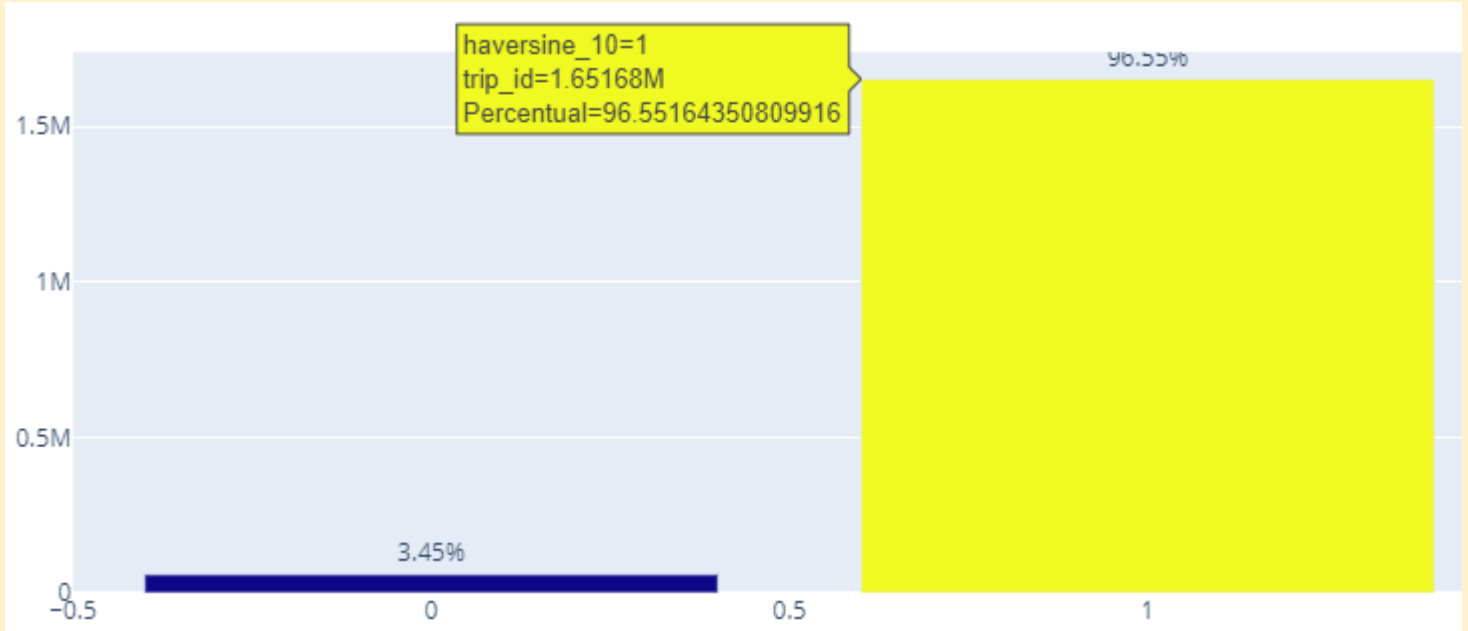
EDA e Hipóteses de negócio

- 6. Ocorrem mais viagens nos meses de alta estação do que no resto do ano. Meses de alta estação: Janeiro (1), Junho (6), Julho(7) e Dezembro(12). Hipótese falsa.
  - Afim de aumentar a lucratividade, a empresa pode fazer parcerias com hoteis e agências de turismo tanto em Portugal quanto em Porto para buscar e deixar turistas. Qual a complexidade para tal?



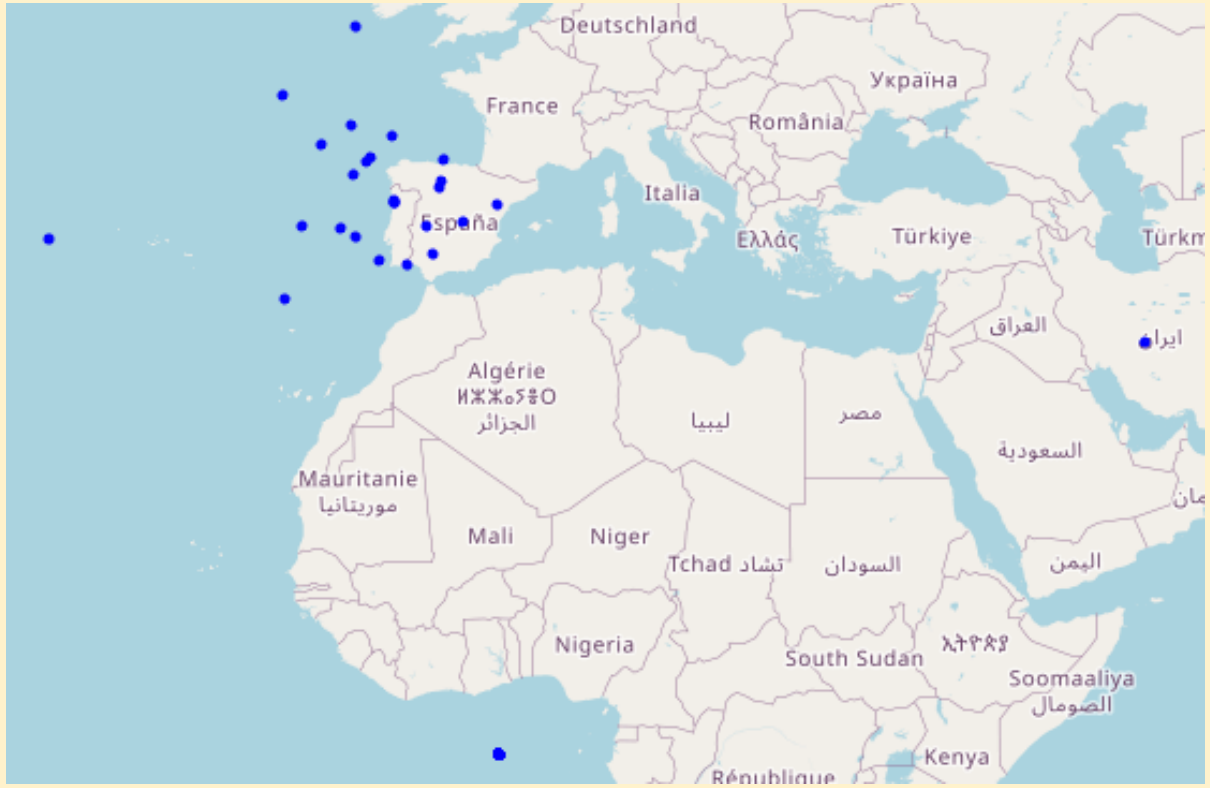
EDA e Hipóteses de negócio

- 7. Pelo menos 70% das viagens são menores que 10km. Hipótese Verdadeira.
- O total de 96,55% das viagens são menores do que 10km. Isso pode indicar também que o sistema de captura de lat e long pode está com algum problema. Devemos levar ao negócio os seguintes questionamentos:
  - 1. O sistema de táxi aceita corrida até quantos quilometros além da cidade de Porto?
  - 2. Caso possa deixar pessoas além da cidade de Porto, como funciona a métrica CALL_TYPE "C", para passageiros pegues em qualquer local fora da cidade de Porto? Os taxistas tem autorização para fazer essas viagens? ou eles só podem pegar as viagens que irão em direção a Porto?
  - 3. Pode ser adicionado aos taxistas de um localizador GPS com chip GSM, para melhorar a captura de lat e long?



Problema do Lat e Long e filtro principal

- A medição Lat e Long possui erros na métrica, sendo necessários retirar dos dados.
- Volto a ressaltar a possibilidade de colocar um rastreador GPS com chip GSM.



Problema do Lat e Long e filtro principal

- Foram realizados 2 recortes distintos para facilitar o entendimento de lat e long.
- O primeiro foi pelo valor absoluto de lat e long inicial e final considerando a dimensionalidade fornecida pelo google maps.
- O segundo foi pela variação (delta) de lat e long inicial e final. Considerando os valores do calculo de haversine e a distribuição normal dos dados, os cortes estão adequados, mas no próximo ciclo do CRISP deve-se reavaliar os cálculos e decisões perante o lat e long conforme modelo de negócio.

- Limitação de latitude e longitude conforme pesquisas e google.

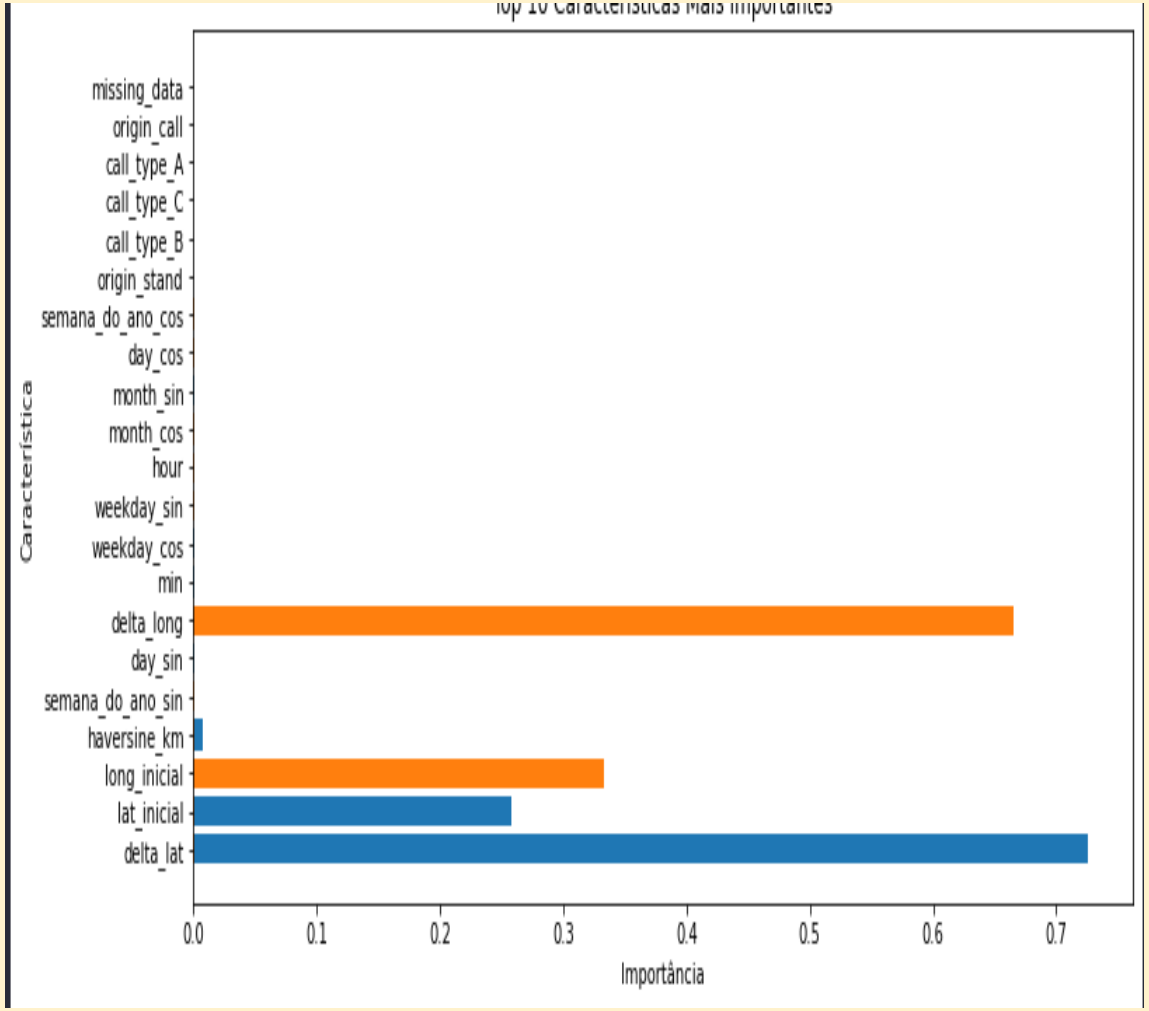
latitude	longitude
42.2	-9.5
38.4	-6.0

- Limitação da variação de latitude e longitude.

latitude	longitude
0.2	-0.2
0.2	-0.2

Preparação dos dados e Seleção dos atributos

- Foi aplicado encoding na coluna call_type.
- Rescaling nas colunas year, haversine_km, delta_lat e delta_long.
  - Teste em outras colunas, mas optei por manter somente nessas 4 devido a distribuição dos dados.
- As colunas relacionado a data apliquei seno e cosseno transformando em variáveis cíclicas.
- Para selecionar as Features, inicialmente tentei com o Boruta, mas devido a limitação computacional não funcionou, então utilizei o método com Random Forest Regressor.
- Após a seleção, de 21 features permaneci com 5 ao total, sendo:
  - delta_lat, lat_inicial, long_inicial, haversine_km, delta_long



Aplicação de Machine Learning

- Como a base de dados era suficientemente grande, optei por utilizar o Método de Validação Houldout aonde eu derivo 3 dataset diferentes do dataset de treino, aonde é possível treinar, validar e testar os dados antes mesmo com maior precisão
- Testei ao total com 6 modelos de regressão, mas com uma peculiaridade. Como eu tinha que prever 2 valores, lat e long final, eu utilizei o método MultiOuputRegressor vindo da bibliote Sklearn no qual me permite fazer essa predição dupla de lat e long.
- Resolvi não utilizar técnicas de cross-validation, nem de finetunning neste primeiro momento. Optei por causa da simplicidade dos dados e posso no próximo ciclo do CRISP implementar essas técnicas.

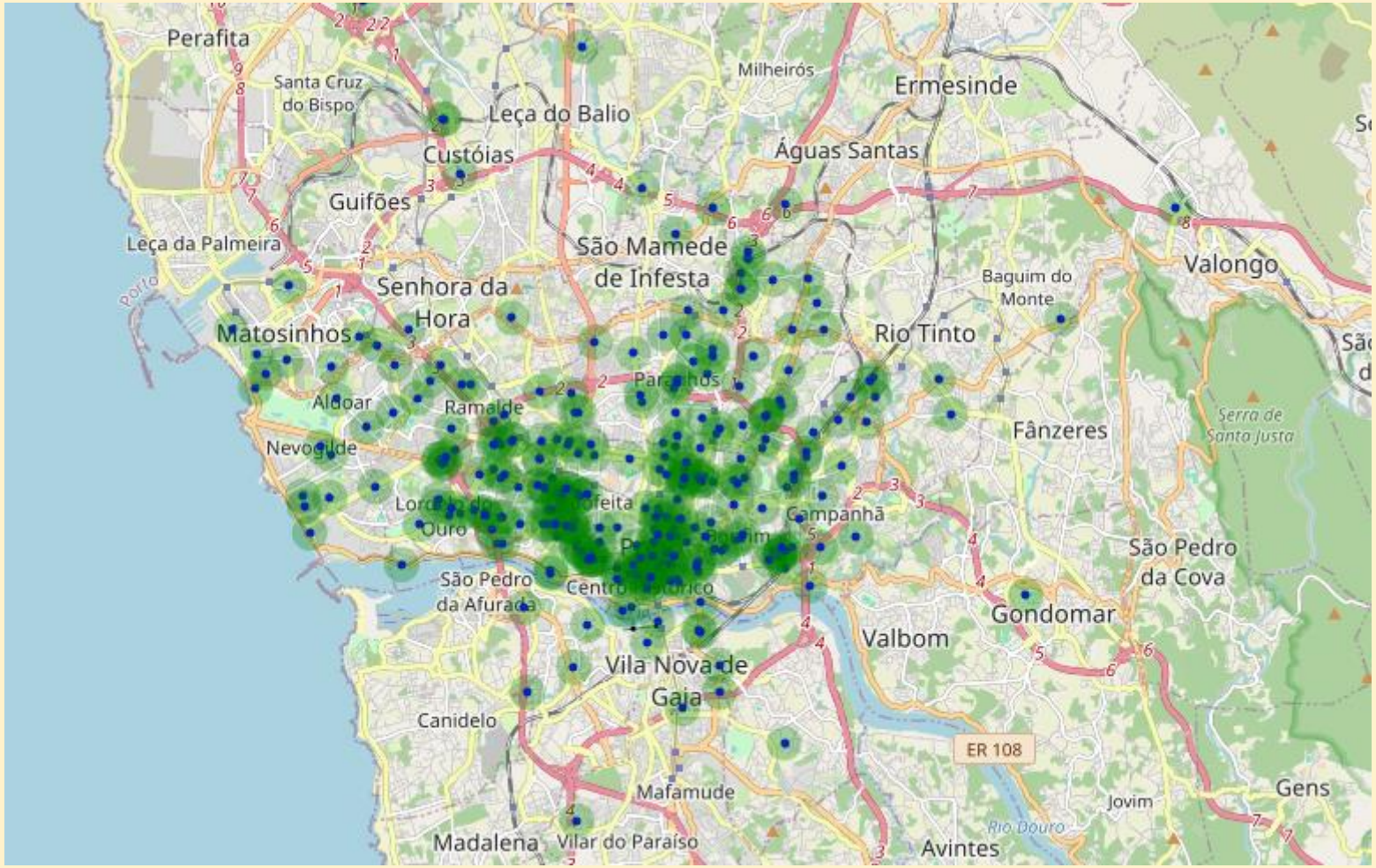
Resultado da aplicação

- Utilizei como métricas de avaliação o R2_score e o RMSE. De forma geral o:
  - R2_score = Mais próximo a 1 representa o quão bem minhas features conseguem explicar o modelo.
  - RMSE = Consigo visualizar a variação da métrica para mais ou para menos na mesma escala. Quanto mais próximo do 0 melhor. O RMSE representa a variação para mais ou para menos do valor central.
- Mantenho minha ressalva de que este modelo em produção é somente para teste e não deve ser utilizado para gerar resultados ao negócio. A segunda versão dele, será lançado em breve com as devidas correções e a possibilidade da equipe de negócio responsável utiliza-lo.

Modelo	R2	RMSE
LinearRegression	0.933698	0.007747

Visualização em Produção

- Com o modelo em produção, o operador poderá ver a cidade de Porto com as possíveis paradas dos taxistas.
- Poderá interagir com o mapa selecionando somente o taxista que deseja, a previsão final e o raio da circunferência ao redor do taxista.



Links de Acesso

- Linked-in: https://www.linkedin.com/in/gabriel-nobre-galvao/
- Portfólio de Projetos: https://bit.ly/portfolio-gabriel-galvao
- GITHUB do projeto: https://github.com/Gabrielnbr/taxi_driver
- API: https://taxi-4ui9.onrender.com
- PROJETO EM PRODUÇÃO: Ainda irei disponibilizar a primeira versão estou resolvendo problemas de compatibilidade