Project Charter

Link to Git Repo

Personnel

Itamar Fradkin (312531064)
Ron Boxer (321219008)

Business background

It is aimed at designing and implementing a basic machine learning pipeline that will contain an automated step that will improve the overall performance of model using any method that doesn't involve changing its type or hyperparameters or isn't dataset-specific.

As a result of the model we describe, real estate agents will be able to provide better estimations of house prices for their clients in Boston.
Moreover, the same model should help to predict how often a driver will file an insurance claim in a year for the French motor claims client. This will allow them to make better pricing.

Scope

One of the first tasks we are taking as we build the MLOps pipeline is to feature engineer our dataset. This includes creating new features by applying polynomial transformations to the original features. This can enhance the performance of our model by capturing non-linear correlations between the characteristics and the target variable. After creating the polynomial features, we are using feature selection techniques to select a subset of the most informative features from the dataset. Specifically, we are using the feature importance provided by the XGBoost algorithm and selecting only the top K features based on their importance scores. This method of feature selection allows us to select the most important features and improve the performance of our model while reducing the overfitting.

Once we are satisfied with the performance of our model, we are packaging it and deploying it using the Streamlit.io platform. This platform allows us to easily build, test and deploy a machine learning model as a web application.

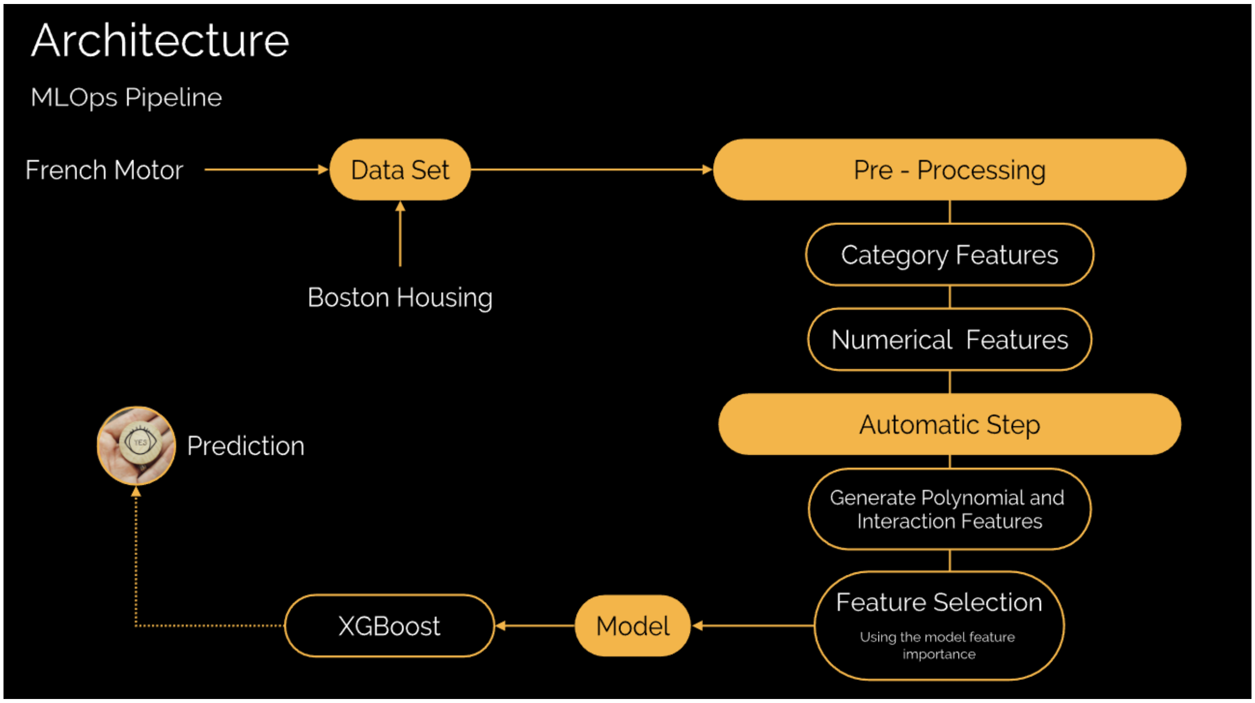
Metrics

We will present the MSE metric on our two data set. When using the MSE on the French motor claims dataset and the Boston Housing dataset, it could be appropriate because both of these datasets are regression problems, where the goal is to predict a continuous variable. In the case of the French motor claims dataset, the goal could be to predict the total amount of claims for each policy holder and in the case of the Boston Housing dataset, the goal could be to predict the value of homes.

Plan

Phase 2 - Set the enviorment. 1 Day
Phase 2 - Train XGboost models on each one of the datasets. 1 Day
Phase 3 - Insert our automated step. 3 Days
Phase 4 - Run locally full pipeline, on each one of the datasets. 3 Days
Phase 5 - Upload the pipleline into a visualization app. 2 Days

Architecture



Architecture

MLOps Pipeline

Datasets-

Boston House Prices Data Set -

There are 606 records. each record in the database describes a Boston suburb or town. The data was drawn from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. The attributes are defined as follows (taken from the UCI Machine Learning Repository1): CRIM: per capita crime rate by town

- ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10 million)
  1https://archive.ics.uci.edu/ml/datasets/Housing
  123
  20.2. Load the Dataset 124
- RM: average number of rooms per dwelling
- AGE: proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centers
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per $10,000
- PTRATIO: pupil-teacher ratio by town 12. B: 1000(Bk−0.63)2 where Bk is the proportion of blacks by town 13. LSTAT: % lower status of the population
- MEDV: Median value of owner-occupied homes in $1000s
  We can see that the input attributes have a mixture of units.

French Motor Claims Dataset -

Contains 6M records and 11 columns -

- IDpol The policy ID (used to link with the claims dataset).
- ClaimNb Number of claims during the exposure period.
- Exposure The exposure period.
- Area The area code.
- VehPower The power of the car (ordered categorical).
- VehAge The vehicle age, in years.
- DrivAge The driver age, in years (in France, people can drive a car at 18).
- BonusMalus Bonus/malus, between 50 and 350: <100 means bonus, >100 means malus in France.
- VehBrand The car brand (unknown categories).
- VehGas The car gas, Diesel or regular.

- Density The density of inhabitants (number of inhabitants per km2) in the city the driver of the car lives in.
- Region The policy regions in France (based on a standard French classification)

Communication-

Our initial communication will be via Email, although in the feature we will add a chatbot you our streamlit platform. This chatbot will be connect to our customer seccsus and will provide any demands from our clients.