Project Charter

Exercise grade: 80

Personnel

Itamar Fradkin (312531064) - Data Scientist
Ron Boxer (321219008) - Data Sceintist

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.

Scope

Using "Feature-Engine"[1] library we will automate the **feature selection process**. We will test different feature extraction methods, including "DropDuplicateFeatures," "DropCorrelatedFeatures" and "SelectBySingleFeaturePerformance" all of which are a part of the package, and we will add a feature selection step that will only use the top K features. In an effort to raise the baseline model's RMSE measure and more.
We want the costumer get an online app that visualize all the process(such as streamlit.io)

Metrics

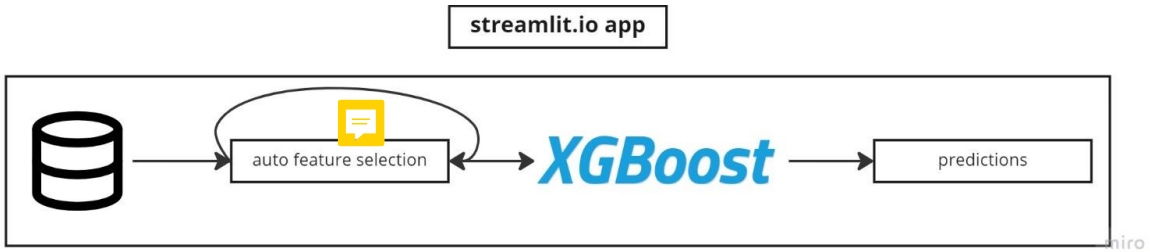
We will present several metrics on our two data sets. MSE, RMSE, MAE, R-squared Score.

Plan

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

[1] https://feature-engine.readthedocs.io/en/1.1.x/

Architecture
Pipeline Scheme-



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)