Mid-Term Exercise

January 23, 2023

1 Business background

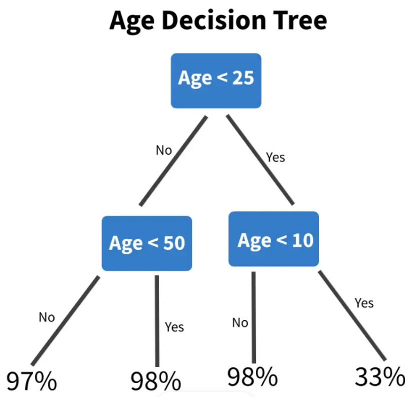
As an ML consultancy group we enhance our customer's AI capabilities by applying our proprietary combination of ML Ops methodologies to improve their ML models. Our current client is a real estate investment firm specialized in locating potential investments in the Boston residential area. The company uses XGBoost ML models in order to predict property prices for these potential investments. The firm seeks to derive more informed business decisions with greater certainty. We will present an enhanced ML model which will enable the firm to hone their modeling capabilities and yield better and more accurate results. Our approach will focus on finding and addressing weaknesses in the given data-set by locating subsets of the data where uncertainty is high.

From a data science perspective this means increasing the accuracy of the baseline model. A binary thresholding process is completed to turn this regression problem into a type of classification problem wherein discussing accuracy becomes relevant. So model predictions which are within a certain threshold of the target value are considered accurate, and predictions outside of this thresholding window are considered inaccurate. Increasing this KPI from a data science perspective connects to the business KPI of decreasing the number of inaccurate predictions, and increasing the overall number of satisfied clients.

2 Scope

In this project, we will utilize several Pre-Processing methods including data normalization, standardization and one-hot encoding. We will harness third party tools to identify weak segments in the data and as a next step will detect uncertainty in certain features. Finally, we will generate synthetic data in order to improve these deficiencies by enriching the training set. All of this will be handled by a sklearn pipeline.

- **FreaAI:** We will implement feature weakness detection algorithm based on IBM's FreaAI tool [2] to find weaknesses in the data. We will use either Highest Density Regions (HDR) or decision trees to find bivariate data slices with low accuracy. These bivariate data slices reduce the search space for categorical predictors and second-order interactions to show where our data is more likely to be problematic. It's worth noting that this tool and its implementation using decision trees is meant for classification problems. That is why our binary thresholding step is crucial for transforming this regression problem into a classification one.



As we can see in the above toy example, by using Decision Trees one can identify slices of data where the base model had low performance.

- **UQ360:** If time permits, we will locate segments in the data with high uncertainty using IBM's UQ360 tool. First we will measure the uncertainty of our baseline model based on the most important features found by the model. To assess uncertainty we will utilize two metrics. The first one is called Prediction Interval Coverage Probability (PICP), included in UQ360. Formally, PCIP is defined as the fraction of a sample (e.g., train data) covered by the prediction interval. The second metric is called Mean Prediction Interval Width (MPIW) which Computes the average width of the the prediction intervals [1]. Consequently, we will re-calibrate the prediction interval using Uncertainty Characteristic Curve to improve the PICP score. By using the UCC Recalibration tool we will try to optimize the PICP/MPIW operating point. this will be implemented at later stage.

After finding the biggest slice of data with the most error in our model prediction (low entropy, high number of samples), we will use YData Synthetic — a package to generate synthetic tabular data — to generate syntactic data and retrain our model to reduce loss.

Once our model is trained, the model and its learned parameters will be exported and the customer will be given a compiled python file that can be executed in single line CLI command with a CSV file, containing house samples to predict, given as a parameter. The program will print predictions based on these samples.

3 Personnel

- Client - Real Estate Brokers in Boston

- Client Point of Contact - Product Team
- Client Business Stakeholders - Product Team

4 Metrics

- Qualitative Objectives:
  - Reduce the number of houses sold significantly under or over their market price.
  - Increase certainty in decision making regarding real estate prospects
- Quantifiable Metric:
  - Reduce the baseline MAE by 20% which currently is 0.95 for the French Motor Claims and which is 2.21 for the Boston House Price predictions

  - Increase the PCIP to be 95% for the most important features based on the baseline model (which is currently 89% and 87% for the Boston House Price predictions and for the French Motor Claims, respectively). the UQ360 related metrics will be implemented at later stage.

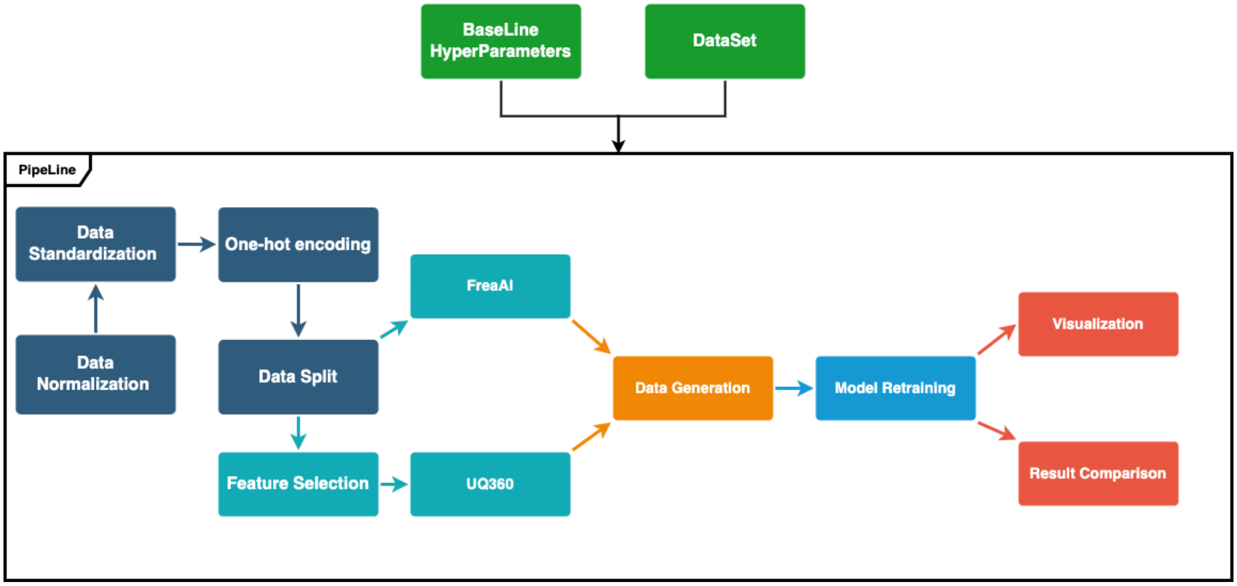
- Metric Measurements - For MAE we compare results of the baseline model to those produced by our enhanced model. For PCIP we will measure the metric of the baseline on the most important features and the compare it to the results achieved on the same features using our enhanced model. the UQ360 related metrics will be implemented at later stage.

5 Plan

- Build a generic Pipeline infrastructure.
- Produce uncertainty metrics for most important features for the baseline model.
- Create Normalization, Standardization, and One-hot encoding steps
- Create and test FreaAI by the following steps:
  - Use Highest Prior Density (HPD) methods to find univariate data slices with low ac- curacy. These univariate data slices reduce the search space and show where our data is more likely to be problematic.
  - Use decisions trees to find bivariate data slices with low accuracy. These bivariate data slices reduce the search space for categorical predictors and second-order interactions to show where our data are more likely to be problematic.
- If time permits, create and test UQ360 uncertainty step
- Create and test syntactic data generation step
- Visualize results
- Run and Debug the entire Pipeline process
- Wrapping everything with Python in order to create a single line CLI command for pipeline usage
- Assemble the results and create presentation

6 Architecture

- Our Project will be created using Python in Google Colab.
- Our data sets are stored in CSV files which will be upload to the RAM using Pandas dataframes.
- We will create a generic pipeline, not dependent on any specific data set, and which can be run using different Hyper-parameters set from different baseline models. In order to do so, we will create a pipeline which receives a Hyper-parameter set and data set as input parameters. At that point, all other steps will be generic and will run the same on any given data set. The pipeline will be built using sklearn pipeline modular package.
- All the Pre-Proccessing steps will be created using packages from sklearn and models from the feature engine library.
- In order to save time while fitting the model using UQ360 and FreaAI, we will use threads, running the training simultaneously.
- In order to make the predictive tool accessible to the costumer, we will create a python file that can be executed as single line CLI. The file will be given a CSV file, containing samples to predict, as a parameter and it will print the results and store them in a separate CSV file.



Some feature selection methods we would use include PCA in order to return a lower dimensional representation of the data while still capturing the most variance and information within the data set. Depending on model results and to preserve interpretability, LDA may be used in place of PCA to reduce dimensionality as a feature selection method.

7 Communication

- A weekly meeting will be conducted between our project manager and the domain expert representative from the real estate firm.
  - In these meetings we will assign action items for both parties, answer any questions team members may have, and review notes from previous meetings to ensure we are on track to finish the project on time.
- Once the model is prepared, our project manager will meet with the firm's IT member for deployment arrangements.

References

[1] Matthew Arnold Jiri Navratil Benjamin Elder. "Uncertainty Characteristics Curves: A Systematic Assessment of Prediction Intervals". In: ().

[2] Marcel Zalmanovici Samuel Ackerman Orna Raz. "FreaAI: Automated extraction of data slices to test machine learning models". In: ().