Exit Report

January 23, 2023

1 Overview

In this report we showcase a modeling scheme used to improve results of a baseline XGBoost model on the Boston Housing Prices and French Motor Claims datasets.

The machine learning pipeline consists of several generic steps meant to improve XGBoost model performance by appropriately pre-processing the input data and identifying weak segments which account for poor predictions within the model.

The ultimate goal is to reduce inaccurate predictions made by the model which can increase customer satisfaction, saving time and money for prospective clients.

2 Business Domain

The Boston based residential real estate company is involved in buying, selling, and managing properties in its surrounding suburban areas. The company uses historical data to analyze various aspects of the housing market such as median property values, trends in housing prices, and demand for different types of properties. This information is used to focus their business strategy by identifying, for example, areas where there is high demand for specific types of properties. This allows the company to make data-driven decisions on which properties to purchase or develop.

The Paris based automobile insurance company is involved in selling and managing car insurance policies for their customers. The company uses historical data to analyze various aspects of the auto insurance market in Paris, such as the frequency and severity of claims, the types of vehicles involved in accidents, and the demographics of the policyholders. This information is used to inform the company's business strategy such as pricing policies, identifying high-risk areas or groups of drivers, or making underwriting decisions. Furthermore they study patterns and trends in claims data which helps them identify areas for improvement in their claims handling process, and develop new products or services. They even take actions to mitigate the risk of fraud.

3 Business Problem

For the Boston based residential real estate company, incorrectly predicting the value of homes may cause a loss of money in several ways:

- Selling a house at a lower price than the true market value
- Buying a house at a higher price than the true market value

Predicting each house by an appraiser is very expensive (salary for all employees) and very inefficient (requires time for transport and visiting each house). Therefore the company wants to create an AI based algorithm to predict houses more precisely, more efficiently, and for less money.

For the Paris based automobile insurance company, incorrectly predicting the number of claims of a client per year will yield inaccurate policy pricing which causes loss of money (for under-priced policies) or loss of clients (for overpriced policies). As in the Boston real estate company, predicting the policy price by a human expert is much more expensive and takes much more time.

4 Data Processing

After analyzing our data, we used several pre-processing steps to improve our model performance:

- **Dealing with categorical and non numeric features**:

  First, we implemented the One-Hot Encoding method for those features. We thought this method would work best because it doesn't assume anything about the different categories "distance". After taking into consideration our next steps (specifically FreaAI) we decided to change our approach and deal with this feature by Ordinal Encoding. By this method, we preserve the number of features in data and the amount of information in each feature.

- **Defining and calculating accuracy for FreaAI Decision Trees:** In order to apply FreaAI decision trees to our data we had to convert our problem from a regression task to a classification task. Therefore, as part of our preprocessing steps, we run a baseline model on our data and tagged each instance as "Accurate" if the prediction of the baseline model was within a threshold from the true value and "Inaccurate" if the prediction wasn't within that threshold.

- **Splitting data to low and high-quality data:** Using FreaAi Decision Trees method we split our data into two groups:

  – "High" quality data - data which the baseline model got a high percent of accurate predictions.

  – "Low" quality data - data which the baseline model got a low percent of accurate predictions.

5 Modeling, Validation

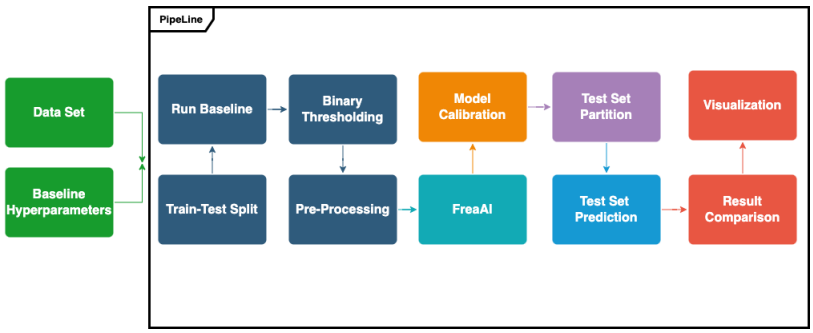
The XGBoost Regressor was the only model used in this workflow, but it was applied in two separate scenarios. Once in order to establish baseline performance and another to validate on FreaAI's output of underperforming data slices. In this case we focused on improving the underperforming data slice's accuracy while maintaining the overall performance.

6 Solution Architecture

Our solution is based on a generalized pipeline that is model-agnostic and data driven which re-trains the existing baseline model. The model pre-processes the data and utilize FreaAI tool to isolating underperforming data via FreaAI tool. Removing underperforming training data yielded the best test results and was implemented into the pipeline. Please refer to the model report for further details. Final output:

- New model equipped for handling new data. The new model is conditioned to partition the data and predict accordingly.

- Interpretability of the model - if requested the customer can seek the exact subgroup in feature space that is partitioned as underperforming. This can be used as a business tool for focusing on specific market sections and avoid treading in dangerous waters.



7 Benefits

- **Boston Real Estate Company:**

  The benefits of using this modeling scheme to predict housing prices include an improvement in the accuracy of housing price predictions, which in turn leads to better decision-making and higher profits. Additionally, this is a powerful and efficient algorithm that can handle large datasets which will save the company time and resources. The customer can also use this to improve its marketing strategies, as he can target specific areas. Overall, the ROI of using this modeling scheme would be significant as it could help the customer make more informed and profitable decisions.

- **French Automobile Insurance Company:**

  The benefits of using this modeling scheme to predict insurance claims include an improvement in the accuracy of insurance claim predictions, which in turn leads to better risk management and lower costs for the customer. Additionally, this is a powerful and efficient algorithm that can handle large datasets which will save the company time and resources. The insurance company can also use this to improve their pricing strategies and potentially attract more customers by pricing their policies more competitively. Overall, the ROI of using this modeling scheme would be significant as it could help the company manage risk more effectively and increase profits by reducing costs and attracting more customers.

8 Learnings

- **Project Execution**

  Customer engagement is crucial for the successful execution of any ML project. It is important to understand the customer's needs and goals, as well as their level of technical understanding, in order to effectively communicate and collaborate throughout the project. Regular check-ins and updates can help ensure that the project stays on track and meets the customer's expectations. It is important to mention that the pipeline is prone to data shift, thus there should be periodical data updates by the customer in order for the model to be kept to date. Additionally, involving the customer in the model selection and evaluation process can help build trust and buy-in for the final product. It's also important to consider the customer's data privacy and security concerns. Overall, effective customer engagement is key for the successful execution of a machine learning project.

- **Data Science/Engineering**

  From a data science standpoint, we conducted a detailed examination of the workings of the XGBoost algorithm to enhance our comprehension of the tool we created. Although our team possessed prior knowledge of the various scenarios in which this algorithm is employed, we desired to acquire a complete understanding of its functions. During our research we got familiarized with the FreaAI algorithm, and were able to utilize it in our specific problem. Along the way we experimented with other tools, such as UQ360 - an uncertainty measuring tool and Ydata Synthetic CWGAN-GP synthesizer in order to generate new synthetic data. From the engineering perspective, as data science students we are used to working mainly with Jupyter notebook and indeed our project was first written in a notebook. At the end, we learned how to convert the notebook into a python script while trying to be pythonic and uphold SOLID principles, and how to work with git to manage the code as a team.

- **Domain**

  In order to better understand the features of each of the datasets we had to do online research. It was especially relevant to the French dataset where we needed to understand terms like Bonus Malus and types of Exposures to figure out the logic behind choosing these features when trying to estimate insurance claims. In a similar manner, we researched about certain features in the Boston dataset to sharpen our intuition on how to approach the data.

- **Product**

  It is important to have a good understanding of the products and services being utilized in any ML solution. This includes understanding the capabilities and limitations of the tools and frameworks being used, as well as how they can be configured and optimized for the specific task at hand. In this project we encountered a lot of new tools and were able to sharpen our knowledge in more familiar tools. We learned about the FreaAI model and how it can be useful in identifying weak data segments. We learned how to work with scikit-learn's pipeline module and different types of encoders. We learned how to generate synthetic data, experimenting with various approaches such as SMOTE and GANs.

- **What's unique about this project, specific challenges**

  Some specific challenges we had to overcome in this project was dealing with largely imbalanced data in the French Motor Cars dataset, and a relatively low number of samples in the Boston Housing dataset. To address these issues we tried generating more data using a generative adversarial network (CWGAN-GP). Also, we utilized numerous new tools to create this pipeline and there was a learning curve until we were able to hone their capabilities.

9 Links

GitHub Repository: https://github.com/itayshap/MLOPS_Final_Project

10 Next Steps

Here are several logical next steps for this project:

- **cross validation**

  We should explore the option of using cross validation in our pipeline. It might be beneficial when dealing with small data like in the Boston dataset.

- **Synthetic Data**

  Further exploration of generating syntactic data in order to offset imbalances in the dataset, this is especially relevant in the French Motor Cars dataset. In regards to the Boston dataset, the creation of new data will likely improve model accuracy as it is such a small dataset.

- **HPD**

  In terms of the FreaAI implementation, stacking the highest posterior density (HPD) method on top of our existing decision trees could further improve the generic step responsible for identifying underperforming data slices and in turn improve model results.

- **FreaAI - regression decision trees**

  Further exploration is needed in our implementation of FreaAI using regression decision trees, eliminating the need to create binary classification. Experiments should include using FreaAI to create trees with more than two features to increase the relevance of the data segments on the full data. Furthermore, several experiments — in regrades to hyper-parameters of the regression decision trees — and their affect on the end results should be formulated. In addition, grid search should be used to optimise the MSE threshold and minimum samples cutoff in the leafs.