G2JN Project Charter - Mid Term Exercise

Reichman University
MLOps Course - December 2023

Business background

The problem and solution this project handles is designing and implementing a machine learning automatic pipeline that improves the performance of a given model. The implementation will handle tabular data and will be suitable to improve the performance of an regression model with XGBoost Regressor as baseline. This pipeline will help customers get better understanding and predictions of their data in order to be more competitive and get better results on their business metrics.

As design partners for this projects, we already have two companies that are engaged and interested in our proposal.

1. A **French Insurance Company** that underwrites motor insurance policies They currently have a dataset of more than 600.000 policies and their claims within a year. It also has different risk features such as geographical on where the vehicle is used as well as drivers information.
2. An uprising **Real Estate Startup** that is expanding its operations to Boston, USA. They are performing its research in order to be able to price the properties they sell. They are using the well-known dataset for boston housing from Boston Standard Metropolitan Statistical Area, which gives house prices with respect to the house characteristics (squared meters, number of rooms), location (crime in the area, closeness to the river, nitric oxides concentration, etc), within other information.

Both companies are standing on the above mentioned baseline of an XGBoost Regression, performed by their team. However, they are hiring consultants to improve their results.

- For the **French Insurance Company,** their specific request is “how can they reduce the error on prediction of claims” so that they can underwrite better their customers?”. In the end, they need to perform better so that their bottom line profit increases by reducing the amount wrong (underperforming) insurance policies given and being able to price higher the most high risk cases. Also, they want not to overprice so that they do not lose potential customers.
- As for the **Real Estate Startup**, their intention is to price fairly each house so that (1) they continue building reputation in the market that will drive winning market share against its traditional competitors, (2) not underprice to leave profit on the table for them and the house owners and (3) not to overprice so that they do not close or lose a deal.

Scope

The main problem the customers have is that their current baseline is not enough to get the performance and quality of results (predictions) that they need for the business. Given the requests and needs mentioned, we intend to build a pipeline that:

1. Analyzes the data and models to identify slices with high aleatoric error and slices with epistemic error. This step will be done with a tool such as Oracle Macest. If needed we will also consider a library for outlier detection such as PyOD.
2. Correct slices with aleatoric error by performing outlier removal. The reason behind this is we understand that some outliers in the data usually drive to higher aleatoric error as they introduce noise and affect precision on the model.
3. Improve slices with epistemic error by generating synthetic data with libraries such as ~~SMOGN or ydata-synthetic~~ Synthetic Data Vault (SDV). The generated data will improve low performing data slices that will enhance the trainset thus making the re-trained model exposed to all slices with balanced data, considering the original distribution of each of them.
4. Retrain the model with the enhanced dataset.

By correcting epistemic and aleatoric error in the dataset, we expect to reduce the error on prediction, which in the end will lead to the business requirements as stated on previous section.

Our clients will interact in the same way as they did with the baseline they had. This is feeding the model with a csv file that contains new samples (of houses or motors) and will get as response a prediction for each sample.

Personnel

- Consulting Company (us):
  - Gil Ayache - 200358612
  - Gil Zeevi - 203909320
  - Joel Liurner - 346243579
  - Nadav Elyakim - 205702368
- Client:
  - Real Estate Startup:
    - Data Scientist that built the baseline
    - Head of Growth leading the expansion strategy
  - French Insurance Company
    - Data Scientist that built the baseline
    - Head of Risk in charge of the team assessing and approving policies.

Metrics

The project objective is to reduce the error that both companies perform when predicting. For French Insurance Company, this means to improve their certainty on the claims that a driver/vehicle will file in a given year so that they underwrite better. For the Real Estate Startup, they want to provide better (more market fair) prices of the properties they sell so that they don't lose any customer due to overpricing but also don't sell under market place to increase revenues. In the end, for both of them, this will be translated into increased top line revenue.

For our team, this metric will be considered RMSE (root mean square error). It will will be compared against the baseline model received from the customer (original dataset trained with XGBoost Regressor).

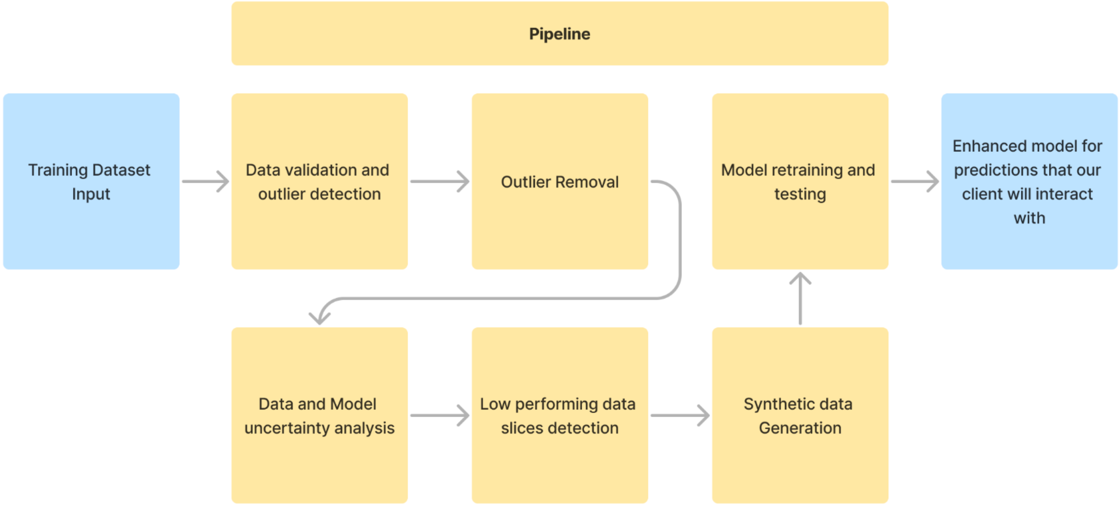
Plan

- 14/12/22 - Start of the project - Get to current baselines.
- 18/12/22 - Implement macest and understand how to assess epistemic and aleatoric error
- 24/12/22 - Get data slices with uncertainty (epistemic and aleatoric).
- 25/12/22 - Remove outliers
- 01/01/23 - Generate synthetic data on low performing data slices.
- 07/01/23 - Automate the pipeline and create unique library or execution file with instructions.
- 08/01/23 - Submit the project.

Architecture

The architecture of the system and process is as follows:

1. Data input from our clients. This will be done with the historical training sets mentioned in the business background section and we will typically accept a .csv file.
2. Step for outlier detection and removal.
3. Step for Uncertainty analysis per data slice and synthetic data generation for its improvement.
4. Regression model training with the enhanced dataset.


Pipeline
Training Dataset Input
Data validation and outlier detection
Outlier Removal
Model retraining and testing
Enhanced model for predictions that our client will interact with
Data and Model uncertainty analysis
Low performing data slices detection
Synthetic data Generation

Once the pipeline is executed on the training data, we get as output a trained model. This model will be the blackbox for the client, in the same way they did have their previous baseline model built by their Data Science team. It will receive a similar csv file with new samples to predict as input (without historic labels), and output another file with the predictions. These final predictions will be useful for the clients to make their decisions.

For Real Estate Startup, the Head of Growth and its team will get a fair price proposal for the property they provide to the model. This will help them to be sure that they continue building their reputation in the market, as well as maximizing its revenues to as much as possible without overpricing their sale.

For the French Insurance Company Head of Risk, he will get a much more accurate prediction on how many claims a certain motor might file, given the characteristics of the vehicle and driver.

Communication

- Before starting the project: we will perform a meeting between our team and both clients' employees involved in the project, to present the milestones and expected work. We will need to align that the metrics used to measure the results are suitable and sufficient for their needs. These meetings will be attended by all stakeholders from the client: the data scientists and head of risk/growth respectively.
- During development: weekly zoom meeting to clarify doubts on the data and show progress. These meetings will be attended by the Data Scientist of our client.
- Product hand-off: final presentation showing a demo of the product and training the relevant teams on how to use it. These meetings will be attended by all stakeholders from the client: the data scientists and head of risk/growth respectively. Their teams will also be invited.