Insurance Claims Modeling Project

Team 18

White Paper

Mentor:

Stefan Ferreira

Introduction

Across the motor insurance industry, organizations are making process efficiency a priority in order to deliver measurable business results, including enhanced profitability and improved customer satisfaction (Maichel-Guggemoos & Wagner, 2018; Alamir *et al.*, 2021). For many companies, whether they provide property and casualty insurance, specialty, reinsurance, or other types, some of the greatest efficiencies can be realized with a focus on the claims process. For example, *InsurTech* reports a 30% increase in customer loyalty for insurance organizations whose customers are satisfied with their claim experience. According to *Accenture*, a mere 2% improvement in loss costs through effective claims management can result in up to $11 billion annual increase in bottom line results.

In recent years, while organizations attempt to improve claims processing, their ability to do so has become increasingly complex for two key reasons. Firstly, the amount of data captured on claims and people, whether they are policyholders or potential policyholders, has increased significantly. Additionally, the variance in policyholders and the non-uniformity of their profiles make predicting their claim severity challenging . While the goal of every insurance company may be clear, often the method for getting there is not. The current standard for claims modeling in the insurance industry has been relatively unchanged for decades, using outdated techniques like generalized linear models, spreadsheets and rule based algorithms to predict claims and set premiums. For a specific company to stand out and outcompete the current insurance landscape, they need to deploy machine learning models, able to evaluate more features and use more sophisticated statistical techniques, while still remaining transparent in their prediction approach.

In this white paper, an overview will be given discussing the data science approach used throughout this internship, where we applied data engineering and machine learning to more accurately predict claim severity.

Problem Statement

Claim estimation is a crucial component of pricing and reserving in the insurance industry. For all insurers, the ability to produce a completely transparent and extremely accurate prediction for a specific risk is essential, as it informs go-to-market strategies and simplifies regulatory reporting. Most importantly, it facilitates accurate pricing by insurers for various risk levels. Over the years, different insurance stakeholders have adopted diverse methods for the estimations of claims. While traditional methods like general linear models (GLMs) are relatively transparent, they lack predictive power and are limited with the amount of features that can be

used, therefore necessitating significant subjectivity in feature management and engineering. On the other hand, deep learning models are typically more accurate, but they need a lot of data, and are limited by their lack of explainability. Considering the above, it is necessary to design a state-of –the-art prediction tool that can boast of high accuracy in all-inclusive claim cost prediction whilst providing full decision transparency.

Design solution

Our team has attempted to solve this problem by studying insurance terms and subject matter. We then proceeded to exploratory data analysis to get a better understanding of underlying correlations and patterns found in the data. For the modeling phase, we started off doing feature selection, feature engineering and data preparation to feed clean data into the model. Following this, we built two machine learning models; A linear regression model and an XGBoost model were trained to fit the particularities of different policyholders commonly found in industry. We used these models to predict claim severity for current policyholders, in order to set premiums for prospective policyholders.

1. Data acquisition

The data used for this project was originally obtained from the client database. However, it was formatted and edited by the team supervisor, Mr. Stefan Ferreira and given to the team in the form of CSV files.

Data reliability and integrity was assured and protected during the acquisition process.

2. Exploratory data analysis

In order to facilitate well-informed model-building, our team conducted EDA on the obtained data. Emphasis was placed on correlation between features and how the features affect claim severity. Several takeaways were obtained from the EDA process. Some of them are outlined below:

- Driver and Vehicle Age: Younger drivers tend to be less careful when driving, increasing the probability of vehicle damage and increased claim severity. On the other hand, due to value depreciation, newer vehicles attract larger claims (Figure 1.1).
- Previous insurer excess: Policy holders were more motivated to claim when previous excess was small (Figure 1.2).

- Vehicle transmission: Since automatic vehicles are generally more expensive than manual vehicles, they were affiliated with larger claim severity (Figure 1.3).
- Marital status: The data indicated that married policyholders hadlower claim amount. This finding was attributed to a larger sense of responsibility -and by extension, lower recklessness tendency (Figure 1.4).

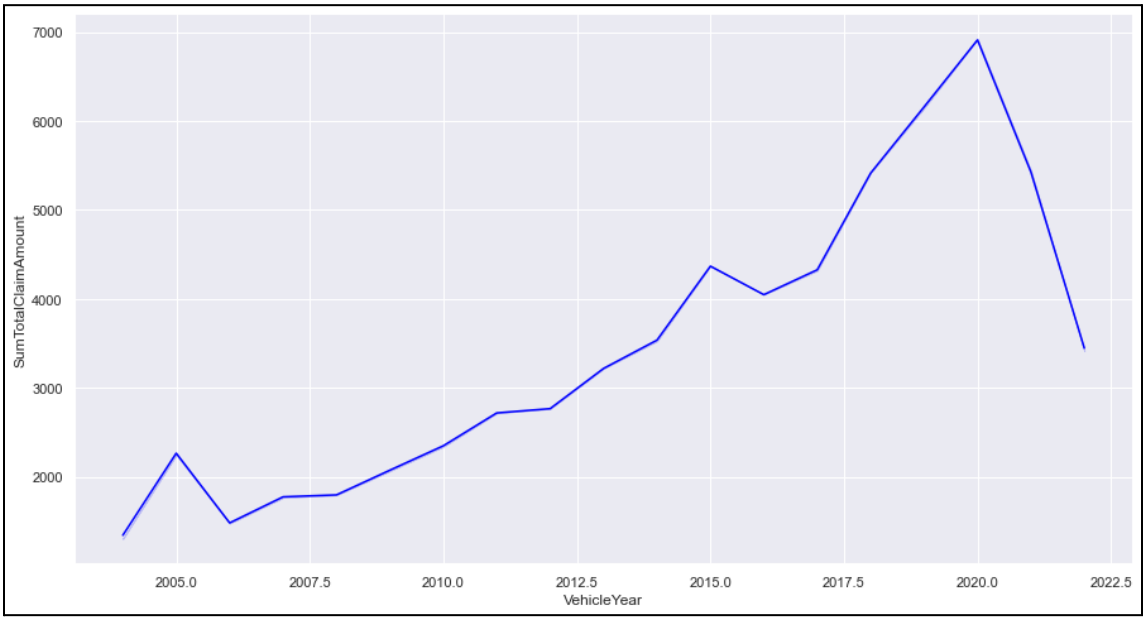

Figure 1.1: Relationship between vehicle age and claim amount

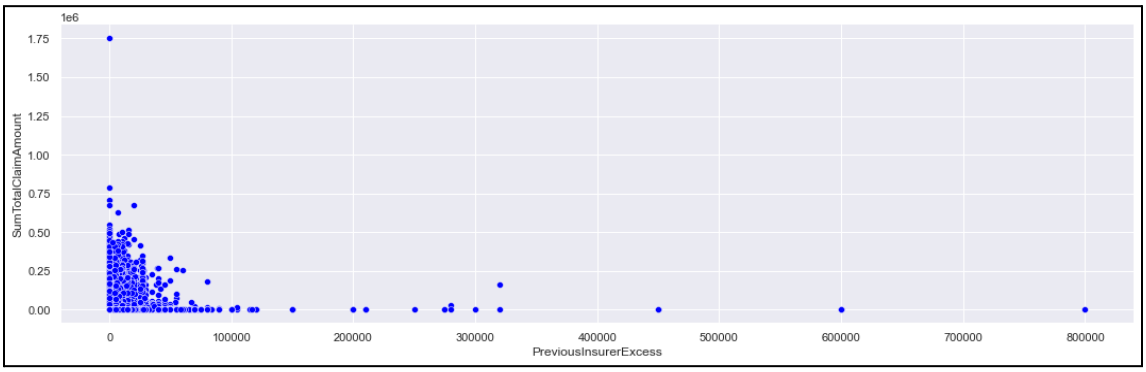

Figure 1.2: Relationship between previous insurer excess and claim amount

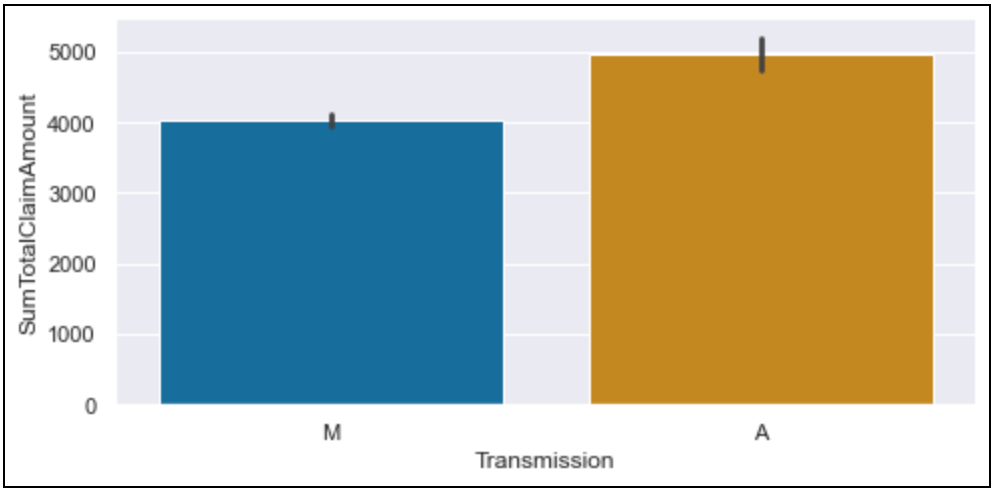

Figure 1.3: Relationship between vehicle transmission and claim amount

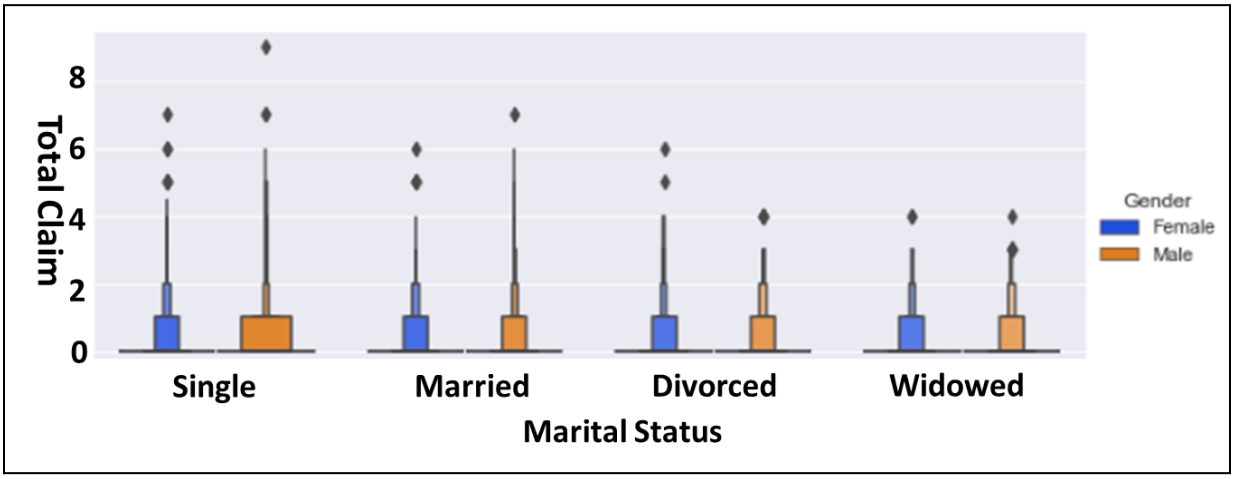

Figure 1.4: Relationship between marital status and claim amount

3. Data cleaning and feature engineering

The team employed the following procedures for the cleaning of the dataset.

- Filling of nulls
- Data type conversion

For the engineering of model features, the team carried out the following procedures:

- Dropping unnecessary columns
- Dropping columns with a high percentage missing values
- Creation of age and exposure features based on other features
- Binarization of numerical features using 5 bins

- Extracting first name of vehicle model
- Selecting top 9 categories and replacing the rest with ‘other’ for categorical features (as this was found to give the best predictions)
- One hot encoding of categorical features
- Scaling of numerical features

For the dropping of features, the criterias for this step were redundancy, feature duplication and high correlation.
After the cleaning and engineering process, out team adopted the following features for model development

The following features were used to train the model
‘PreviousInsurerExcess',‘PreviousInsurerPremium','EmploymentType','IsMemberPayer', 'Occupation', 'IndustryType', 'Gender', 'MaritalStatus', 'Make', 'Model', 'Colour', 'Transmission', 'VehicleType','BodyType','CubicCapacity','Cyl','Kilowatts','VehicleYear', ‘PolicyMainDriverAnnualMileage','PolicyMainDriverLicenseDurationRange','PersonProvince', 'SumAssured','MeanExcess','TotalExcess','ExcessTypesCount','BaseExcess', ‘NomimatedDriversCount', 'NomimatedDriversUnder30Count', 'Exposure', 'Age'.

3.4 Model selection and architecture

For this project, a Linear, then an XGBoost model were trained and used by our team to make claim amount predictions. For the model design, 70% of the dataset was used for training, 20% for validation while 10% was reserved for model testing. To ensure optimal performance, the team carried out hyperparameter tuning using GridSearchCV. The optimal parameters used for the modeling process is shown in Table 1. The learning curve for the XGBoost model is shown in Figure 2.

max_depth	3
min_child_weight	1
gamma	0
colsample_bytree	0.3
max_cat_to_onehot	4

Table 1: Tuned parameters for the teams’s XGBoost model

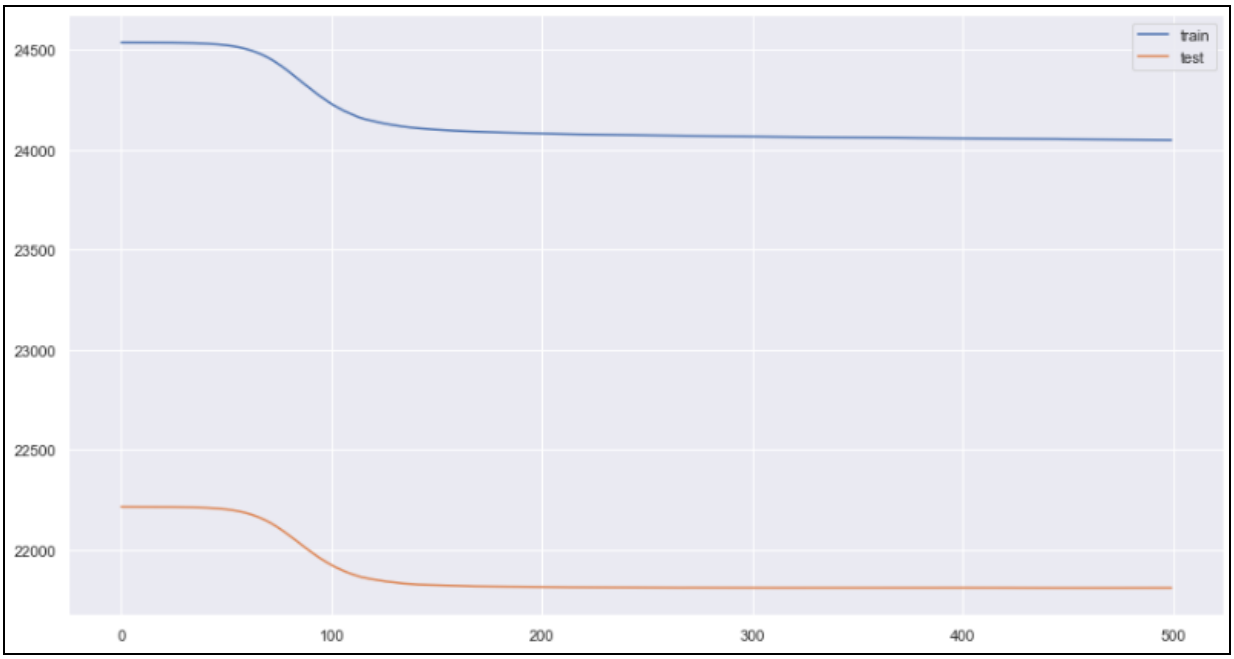

Figure 2: Learning curve for the XGBoost model

The accuracy of the base linear model and XGBoost model predictions were evaluated using the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Explained variance. The findings from the evaluation process is shown in Table 2. Figure 3 is a plot showing the distribution of the actual and predicted claim amount for the XGBoost model

From the result it was observed that the XGBoost produced excellent improvement over the linear regression model. A SHAP plot showing the most important features used by our model for the prediction process is shown in Figure 4.

	MAE	RMSE	Explained Variance
Linear Regression	7416.35	24271.48	0.044
XGBoost	7774.62	23551.36	0.102

Table 2: Result of model performance evaluation

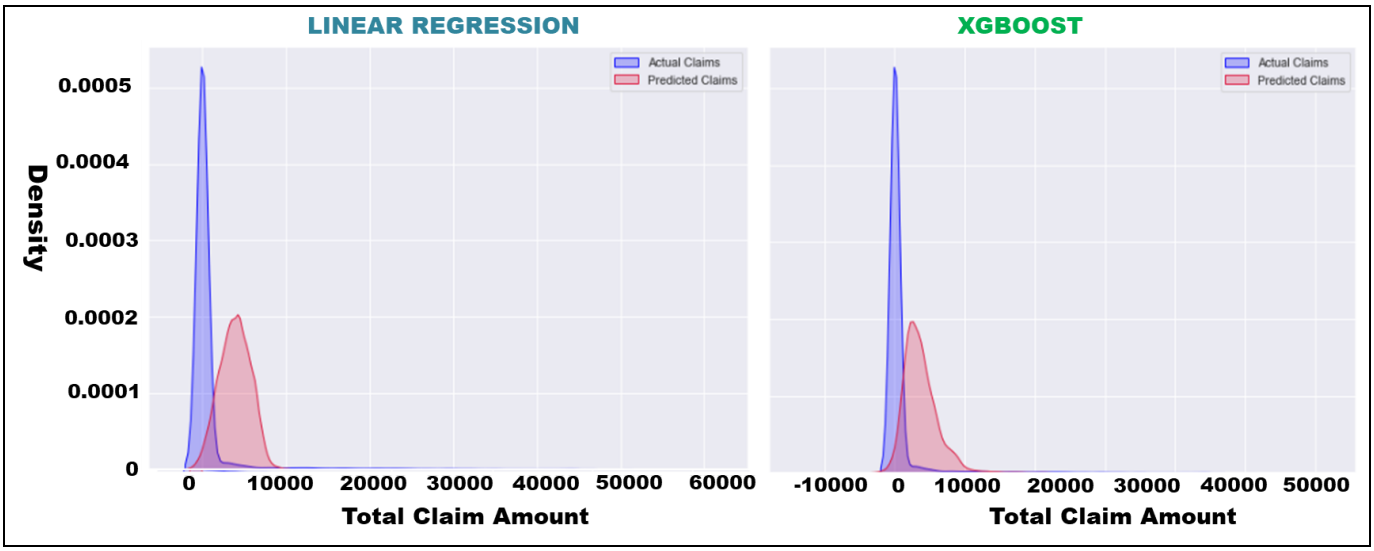

Figure 3: Plot showing the distribution of the actual and predicted claim amount for the Linear and XGBoost models.

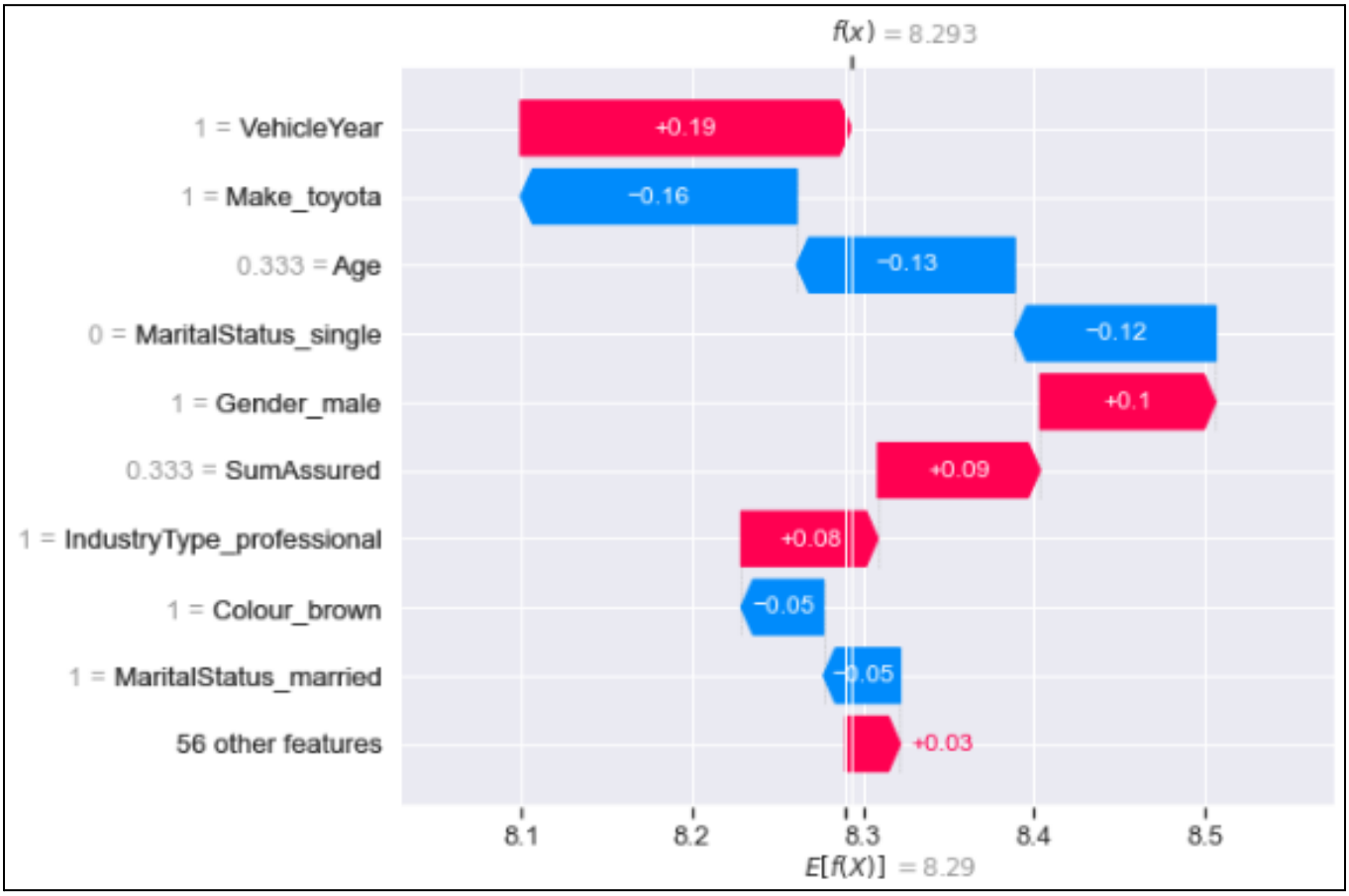

Figure 4: SHAP plot for our teams XGBoost model

4. Model Deployment

The application was developed using Streamlit and deployed remotely on the AWS cloud platform. For remote deployment, the model - built on an AWS EC2 channel was fed by an S3 Bucket where the raw, clean, train and test data were stored.

To facilitate this process, our team constructed and followed the data pipeline which is shown in Figure 5.

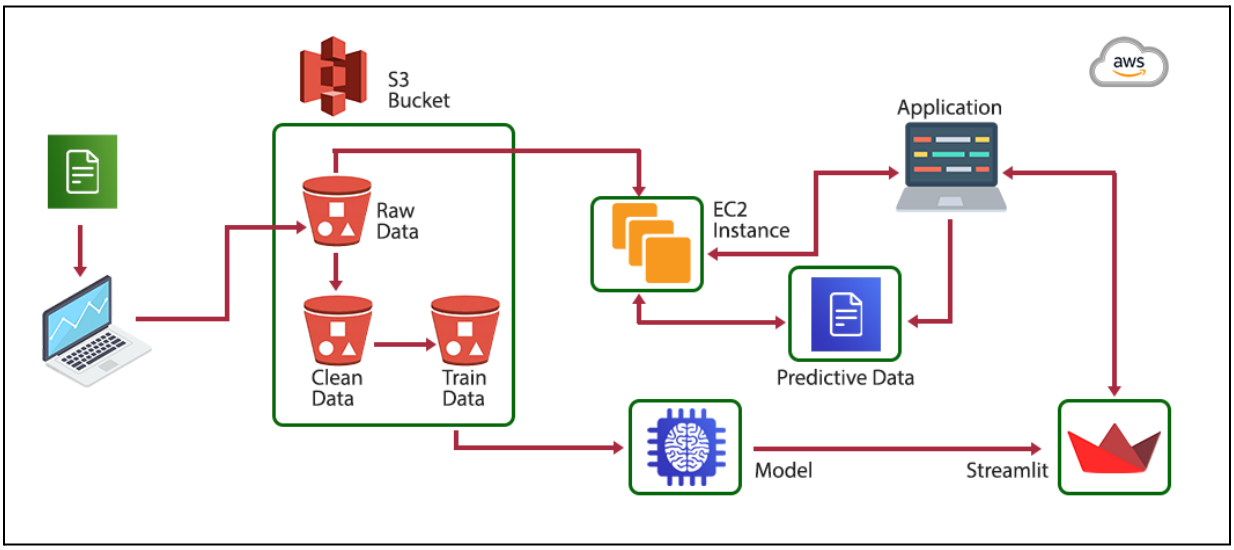

Figure 5: Data Engineering pipeline

5. Conclusion

The current standard for claims modeling in the insurance industry has been relatively unchanged for decades, using outdated techniques like generalized linear models, spreadsheets and rule based algorithms to predict claims and set premiums. These outdated methods are plagued by inaccuracy and poor transparency necessitating search for better models. In this internship, our team employed data engineering and machine learning techniques to develop a robust model for the accurate prediction of claim severity. The tool which was based on the XGBoost ensemble model was found to outperform the traditional Linear regression model presently used by industry. Our team developed the model on streamlit while deploying it on the AWS cloud computing platform.

6. References

Alamir, E., Urgessa, T., Hunegnaw, A., & Gopikrishna, T. (2021). Motor Insurance Claim Status Prediction using Machine Learning Techniques. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/IJACSA.2021.0120354

Maichel-Guggemoos, L., & Wagner, J. (2018). Profitability and Growth in Motor Insurance Business: Empirical Evidence from Germany. *Geneva Papers on Risk and Insurance: Issues and Practice*. https://doi.org/10.1057/s41288-017-0053-4