Insuarance Prediction: EDA and Modelling

Aanal Shah
M.Sc. Applied Modeling and
Quantitative Methods in Big Data
Analytics
Trent Universty
Peterborough, Ontario
aanalmayurbhaishah@trentu.ca

Parth Patel
M.Sc. Applied Modeling and
Quantitative Methods in Big Data
Analytics
Trent University
Peterborough, Ontario
partpatel@trentu.ca

Kavya Gandhi
M.Sc. Applied Modeling and
Quantitative Methods in Big Data
Analytics
Trent University
Peterborough, Ontario
kavyagandhi@trentu.ca

Harshal Panchal
M.Sc. Applied Modeling and
Quantitative Methods in Big Data
Analytics
Trent University
Peterborough, Ontario
harshalpanchal@trentu.ca

***Abstract*— This research employs a dataset of predicting health insurance owners who might interested in Vehicle Insurance, in order to construct a robust predictive model for accurate insurance rate forecasting, with a primary focus on contributing to the field of insurance analytics and aiding providers in optimizing operational procedures. Beginning with a thorough Exploratory Data Analysis (EDA) to unveil inherent features and patterns, the study utilizes visualization techniques to effectively represent correlations and identify crucial factors impacting insurance projections. Subsequently, the project advances to develop a prediction model using advanced machine learning techniques, specifically designed to estimate insurance premiums with a high degree of accuracy. The anticipated outcomes hold significant potential to revolutionize the insurance analytics landscape, providing insurance firms with a valuable tool to refine risk assessment and pricing procedures. The combined power of EDA and machine learning in this project aims to enhance the accuracy of insurance forecasts, offering a pragmatic solution to augment decision-making processes and contribute to the overall advancement of the insurance analytics domain.**

Keywords—Predictive Modelling, Eexploratory data analysis, Feature Selection, Visualizations, Model Evaluation.

I. Introduction

In the dynamic landscape of the insurance industry, the ability to accurately predict insurance premiums stands as a pivotal challenge. The repercussions of imprecise premium estimations extend beyond financial ramifications for policyholders; they also impact the stability of insurance providers. Recognizing the criticality of this issue, our research endeavours to contribute a robust predictive model aimed at revolutionizing how insurance firms assess risk, adjust pricing, and, ultimately, serve the interests of both policyholders and insurers. The financial stability of insurance providers and the satisfaction of policyholders hinge upon the precision of premium forecasts. Inaccuracies in these predictions can lead to financial instability within insurance firms, resulting in potential hardship for policyholders. This initiative is motivated by the imperative realization of the pivotal role accurate premium predictions play in maintaining the equilibrium of the insurance sector. Our research seeks to mitigate these challenges by developing an advanced predictive model using insights derived from the Kaggle dataset.

Our primary objective is to construct a highly accurate predictive model for estimating insurance premiums. This model will utilize diverse algorithms, including Random Forest, Logistic Regression, and K-means clustering, leveraging insights obtained through exploratory data analysis (EDA) and data visualization tools. By identifying crucial variables and correlations in the dataset, our model aims to provide insurance firms with a powerful tool to enhance their risk assessment capabilities and adjust pricing strategies accordingly. Beyond the technical aspects, our research also aspires to offer practical insights for the insurance industry. Through EDA and data visualization techniques, we seek to extract meaningful insights from the dataset, illuminating significant variables that influence insurance rates. The knowledge gained from this analysis can empower insurance providers to refine their risk assessments and optimize pricing strategies, ultimately improving their operational procedures.

The research scope encompasses the comprehensive exploration of the Kaggle dataset, employing advanced modelling techniques and analytical methodologies. We plan to delve into the intricacies of Random Forest, Logistic Regression, and K-means clustering to develop a predictive model of exceptional precision. Additionally, our research will utilize EDA and data visualization tools to unravel insights that can be translated into practical recommendations for the insurance industry. This paper unfolds in subsequent sections, detailing the dataset, methodology, and results obtained from our predictive models. It aims to serve as a valuable resource for insurance professionals, providing them with actionable insights to navigate the challenges of premium estimation and enhance their overall operational efficiency.

II. Previous work

In the vast landscape of insurance analytics, a wealth of research has been dedicated to refining predictive models and enhancing risk assessment methodologies. The evolution of methodologies and the incorporation of advanced techniques have marked key milestones in addressing the nuanced challenges faced by the insurance sector. The following subsections provide a detailed exploration of the key themes in previous work.

A. *Predictive Modeling in Insurance:*

Historically, the application of predictive modeling in insurance has been multifaceted. Traditional statistical methods, such as linear regression, have been employed to model the intricate relationships between various factors influencing insurance premiums. Researchers have delved into understanding the impact of demographic variables, historical claims data, and external economic indicators on the pricing of insurance products. These studies underscore the importance of nuanced modeling techniques for capturing the complexity inherent in insurance datasets.

B. *Machine Learning Applications:*

In recent years, the surge in computational capabilities has paved the way for the integration of machine learning algorithms into insurance analytics. Ensemble methods, notably Random Forest, have gained prominence for their ability to handle nonlinear relationships and capture complex interactions within data. Logistic regression has been widely utilized for binary classification problems, such as predicting customer interest in insurance products. The XGBoost algorithm, known for its efficiency in handling large datasets, has been explored to improve predictive accuracy. These applications showcase the industry's inclination toward leveraging advanced algorithms to enhance the precision of premium predictions.

C. *Exploratory Data Analysis (EDA) and Visualization:*

The pivotal role of Exploratory Data Analysis (EDA) and visualization in extracting meaningful insights from insurance datasets cannot be overstated. Previous work has emphasized the importance of thorough data cleaning processes to handle missing values, outliers, and data quality issues. Univariate analysis techniques, including statistical measures and visualizations, have been employed to gain insights into individual variables. Bivariate and multivariate analyses, such as scatter plots and correlation matrices, have been pivotal in exploring relationships between variables. Feature engineering, an essential component of EDA, involves transforming and engineering features to enhance model performance. These practices collectively contribute to a deeper understanding of the dataset's characteristics and inform subsequent modeling decisions.

D. *Customer Segmentation:*

Tailoring insurance offerings to different customer segments has emerged as a focus area in prior research. Customer segmentation strategies, often facilitated by clustering techniques like K-means clustering, aim to identify distinct groups based on behavior, demographics, and historical interactions with insurance products. By personalizing communication strategies and product offerings, insurers can improve customer engagement and satisfaction.

E. *Gap in Existing Literature:*

While the existing body of literature has contributed substantially to the field of insurance analytics, there exists a notable gap pertaining to the comprehensive exploration of the Kaggle dataset for predicting customer interest in vehicle insurance. Our research aims to bridge this gap by leveraging a diverse set of modeling techniques, incorporating insights from EDA, and providing actionable recommendations tailored to the unique characteristics of the dataset under consideration.

III. METHODOLGY

A. *Data Understanding:*

1) *Features in the Insurance Dataset:*

The insurance dataset under consideration encompasses a diverse set of features that play a pivotal role in predicting insurance outcomes. Table 1 presents a comprehensive list of features along with a brief description of their potential relevance to the prediction task.

Feature Name	**Description**
Id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Driving License	0: Customer does not have DL, 1: Customer already has DL
Region code	Geographic region of the policyholder
Previously Insured	1: Customer already has Vehicle Insurance, 0: Customer doesn't have Vehicle Insurance
…	Additional features specific to the dataset

B. *Data Preprocessing:*

1) *Handling Missing Values:*

The insurance dataset underwent a meticulous process to address missing values, a critical aspect of ensuring the robustness of subsequent analyses. Missing values were identified across various features, and the decision-making process for handling them involved thoughtful consideration of the nature of the missingness and its potential impact on the predictive models.

- Imputation: Missing values in numerical features were imputed using appropriate statistical measures, such as mean or median imputation, to maintain the integrity of the dataset. The rationale behind imputation lies in preserving the overall distribution of the data and preventing the loss of valuable information.

- Removal: Instances with missing values in critical target or predictor variables were carefully considered for removal from the dataset. This decision was made to prevent the introduction of bias in subsequent analyses and model training.

2) Encoding Techniques for Categorical Variables:

Categorical variables, such as 'Region' and 'Smoker,' required specialized encoding techniques to transform them into a format suitable for machine learning models. The chosen method was one-hot encoding, a process that creates binary columns for each category within a categorical variable. This approach was preferred for its ability to maintain the distinctiveness of each category without introducing ordinal relationships, which might be misleading in the context of non-ordinal categorical variables like 'Region' and 'Smoker.

The encoding process was crucial for ensuring that categorical information could be effectively utilized by machine learning algorithms. It aimed to prevent misinterpretation of categorical variables as ordinal or numerical, which could lead to erroneous model predictions.

C. Feature Selection:

1) Rationale for Feature Selection:

The selection of features for the predictive model was a strategic process guided by the desire to enhance model performance, interpretability, and generalization to new data. Features were chosen based on their potential significance in influencing insurance charges, the primary target variable in this analysis. The rationale for selecting specific features is outlined below:

- Relevance to Insurance Charges: Features such as 'Age,' 'BMI,' 'Smoker,' and 'Number of Children' were deemed inherently relevant to insurance charges, given their expected impact on an individual's health and associated healthcare costs.
- Dimensionality Reduction: To mitigate the risk of overfitting and improve computational efficiency, features with limited variation or negligible impact on the target variable were excluded. This dimensionality reduction aimed to retain only the most informative variables for modeling.
- Practical Interpretability: Features were selected to ensure practical interpretability of the model's predictions. This involved prioritizing features that could be easily understood and explained, both from a statistical and business standpoint.

2) Techniques for Feature Selection:

Several techniques were employed to identify and retain the most impactful features for predictive modelling:

- Correlation Analysis: Pairwise correlation analysis was conducted to identify relationships between features and assess multicollinearity. Highly correlated features were carefully evaluated, and decisions were made to retain or exclude them based on their individual contributions to the predictive task.
- Feature Importance from Models: Ensemble learning techniques, such as Random Forest, were leveraged to evaluate feature importance. Features contributing significantly to the reduction of impurity or the improvement of predictive accuracy were given higher importance scores. This approach provided a data-driven perspective on feature relevance.

D. Model Building:

1) Dataset Splitting:

The insurance dataset was judiciously divided into training and testing sets to facilitate robust model evaluation. The commonly adopted 80-20 split was employed, with 80% of the data reserved for training the machine learning models and the remaining 20% allocated for assessing model performance on unseen data. This partitioning strategy aims to strike a balance between providing an adequate amount of data for training and ensuring a sufficient dataset for robust model evaluation.

2) Choice of Machine Learning Algorithms:

The selection of machine learning algorithms was guided by the nature of the insurance prediction task. Given that the goal is to predict insurance charges, a regression problem, several algorithms were considered:

- Linear Regression: Chosen for its simplicity and interpretability, making it a baseline model for comparison.
- Random Forest Regression: An ensemble method capable of capturing complex relationships in the data and handling non-linear patterns.
- Gradient Boosting Regression: Utilized for its ability to build sequentially improving models, thereby boosting predictive accuracy.

3) Training Process and Cross-Validation:

The training process involved fitting each selected model to the training dataset. To ensure the robustness of the models and guard against overfitting, k-fold cross-validation was implemented. Specifically, a 5-fold cross-validation strategy was employed, partitioning the training data into five subsets. The models were trained on four of these subsets and validated on the fifth, with this process repeated five times. This approach allows for a more stable estimation of model performance.

4) Hyperparameter Tuning:

Hyperparameter tuning was conducted to optimize the performance of the machine learning models. Grid search, a systematic approach to hyperparameter optimization, was employed. It involves specifying a range of hyperparameter values and exhaustively searching through all possible combinations to identify the optimal set.

E. Model Evaluation:

1) Evaluation Metrics:

The performance of the predictive models was assessed using a comprehensive set of evaluation metrics tailored to the nature of the insurance prediction task. The following metrics were chosen to provide a holistic understanding of model performance:

- Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual insurance charges, providing a straightforward measure of prediction accuracy.
- Root Mean Squared Error (RMSE): Calculates the square root of the mean of squared differences between predicted and actual charges, emphasizing the impact of larger errors.

2) Visualizations of Model Performance:

To enhance the interpretability of model results, visualizations were employed:

- Actual vs. Predicted Plots: Scatter plots illustrating the relationship between actual and predicted insurance charges, allowing for a visual assessment of model accuracy.
- Residual Plots: Visual representation of the differences between actual and predicted values, aiding in the identification of patterns or trends in prediction errors.
- Model-specific Visualizations: Tailored visualizations such as partial dependence plots or SHAP (Shapley Additive Explanations) values were utilized to interpret the impact of individual features on predictions.

IV. Result

In the dynamic landscape of insurance, understanding and predicting customer behaviour is crucial for companies seeking to optimize their service offerings. This research project delves into the nuanced task of forecasting customer interest in Vehicle Insurance within the client base of an insurance company already providing Health Insurance. The investigation utilizes a robust dataset encompassing diverse demographic factors, vehicle details, and policy-specific information. The overarching goal is to develop predictive models that can discern whether customers, previously engaged with Health Insurance, are likely to express interest in Vehicle Insurance.

This visualization provides a comparison of the distribution of the 'Age' column between the training and testing datasets using histograms. By examining the histograms, we can compare the shape and spread of the age distribution between the training and testing datasets.

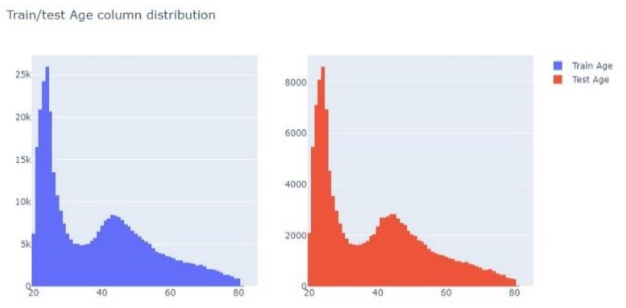

Figure 1 Comparison of train and test Age Variable

The models employed in this study represent a spectrum of methodologies, ranging from traditional statistical methods to advanced machine learning algorithms. A comprehensive assessment of each model's performance was conducted using two key metrics—Accuracy and F1 Score.

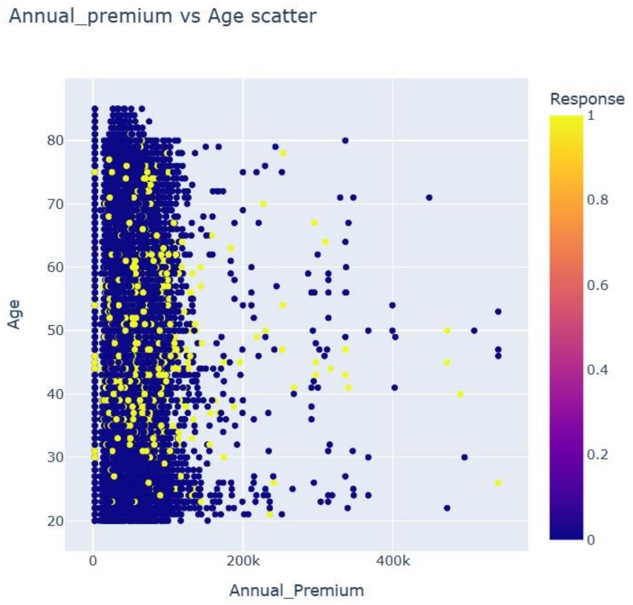

Figure 2 Scatter Plot of Annual Premium Vs Age

The K-Means Clustering model demonstrated a commendable balance between precision and recall, achieving a 74% accuracy and a relatively higher F1 score of 15%. This suggests a moderate accuracy in predicting positive instances, indicating its potential utility in identifying customers interested in Vehicle Insurance. COPOD Anomaly Detection displayed improved accuracy at 76%, although with a lower F1 score of 13%. While surpassing K-Means in accuracy, challenges in accurately identifying positive instances indicate the need for further refinement. Logistic Regression exhibited high accuracy at 87%, but a notably low F1 score of 6% raises concerns about its effectiveness in capturing true positive instances. This prompts considerations regarding potential class distribution imbalance or inherent difficulties in identifying interested customers.

The LightGBM Classifier achieved an 88% accuracy, yet exhibited an extremely low F1 score of 5%, indicating challenges in effectively identifying positive cases. Fine-tuning and optimization are suggested to address this imbalance. The XGBoost Classifier emerged as a standout performer with a balanced 76% accuracy and a high F1 score of 45%. This indicates effective precision and recall, positioning XGBoost as a promising model for predicting customer interest in Vehicle Insurance.

Despite achieving an accuracy of 88%, Random Forest struggled with a zero F1 score, suggesting challenges in identifying positive instances. Further investigation and parameter tuning are recommended for improved performance. The Keras Sequential Model exhibited a tradeoff between overall accuracy and precision in positive instance identification. With an accuracy of 68% and a higher F1 score of 42%, it showcased proficiency in correctly identifying positive cases.

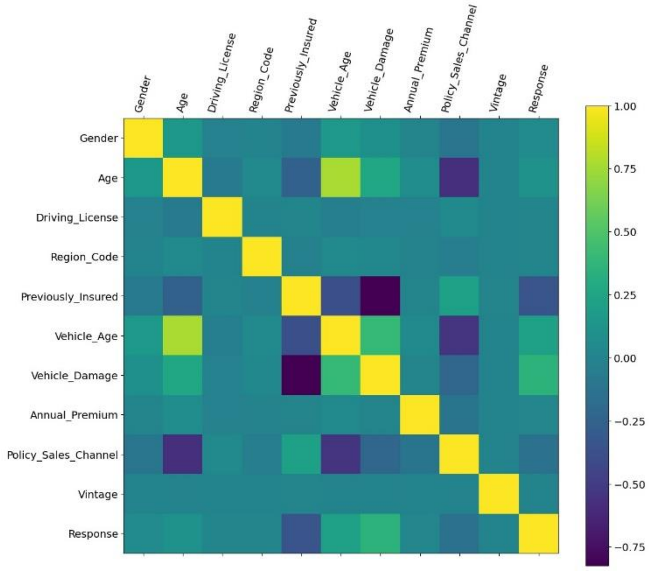

Figure 3 Correlation of Variables

Strong Positive Correlations: Lighter squares indicate strong positive correlations. For example, if 'Vehicle_Damage' is 1, there might be a positive correlation with the 'Response' variable. Strong Negative Correlations: Darker squares indicate strong negative correlations. For example, 'Previously_Insured' and 'Response' have a strong negative correlation, suggesting that customers who were previously insured are less likely to respond positively. Weak Correlations: Gray squares represent weak correlations, and some pairs of features may not show a strong linear relationship.

Model	**Accuracy**	**F1 Score**
K-mean	74%	15%
COPOD Anomaly Detection	76%	13%
Logistic Regression	87%	6%
LightGBM Classifier	88%	5%
XGB Classifier	76%	45%
Random Forest	88%	0%
Keras Sequential Model	68%	42%

V. CONCLUSION

In conclusion, the findings from our extensive exploration into predictive modelling for customer interest in Vehicle Insurance within an existing Health Insurance client base shed light on the intricate interplay of diverse methodologies. While K-Means Clustering demonstrates a commendable balance between precision and recall, XGBoost emerges as a standout performer, showcasing effectiveness in both accuracy and F1 score. These nuanced insights equip the insurance company with a toolkit for tailored communication strategies, enabling targeted engagement with potential Vehicle Insurance customers.

The implications of this study extend beyond the realm of individual model performances. The trade-offs observed between accuracy and F1 score underscore the inherent challenges in balancing precision and recall within a binary classification task. The identified models offer varying degrees of efficacy, prompting further considerations for parameter tuning and model refinement.

Moving forward, the implications of this research extend to strategic decision-making within the insurance industry. The adoption of predictive models, such as XGBoost, can enhance customer targeting and communication strategies. However, the nuanced nature of customer behaviour prediction suggests a need for ongoing research, exploration of hybrid modelling approaches, and continuous adaptation to evolving datasets.

Furthermore, the findings from this study contribute to the broader discourse on insurance analytics, offering insights that can inform industry best practices. As insurance companies navigate an increasingly data-driven landscape, the ability to effectively predict customer behaviour becomes paramount. This study not only provides practical recommendations for the immediate context of Health Insurance cross-selection but also lays a foundation for further advancements in the field of customer-centric predictive analytics.

In practical terms, the implications of this research extend to the optimization of business models and revenue streams. Tailoring communication strategies based on predictive insights allows the insurance company to not only enhance customer engagement but also strategically position itself in a competitive market.

In conclusion, the multifaceted implications of this research underscore the significance of a nuanced and adaptive approach to predictive modelling within the insurance industry. As technology and data analytics continue to evolve, the insights gleaned from this study serve as a stepping stone for insurance companies to navigate the complexities of customer behaviour prediction and, ultimately, optimize their business strategies.

VI. REFERENCES

1. Ejiyi, Chukwuebuka Joseph, et al. "Comparative analysis of building insurance prediction using some machine learning algorithms." (2022).

2. Sahai, Rahul, et al. "Insurance Risk Prediction Using Machine Learning." The International Conference on
3. Data Science and Emerging Technologies. Singapore: Springer Nature Singapore, 2022.

4. Hanafy, Mohamed, and Ruixing Ming. "Machine learning approaches for auto insurance big data." Risks 9.2 (2021): 42.

5. Rukhsar, laiqa, et al. "prediction of insurance fraud detection using machine learning algorithms." mehran university research journal of engineering & technology 41.1 (2022): 33-40.

6. Azzone, Michele, et al. "A machine learning model for lapse prediction in life insurance contracts." Expert Systems with Applications 191 (2022): 116261.

7. Díaz, Zuleyka, et al. "Machine learning and statistical techniques. An application to the prediction of insolvency in Spanish non-life insurance companies." The International Journal of Digital Accounting Research 5.9 (2005): 1-45.

8. Boodhun, Noorhannah, and Manoj Jayabalan. "Risk prediction in life insurance industry using supervised learning algorithms." Complex & Intelligent Systems 4.2 (2018): 145-154.

9. Soloviev, Vladimir, and Vadim Feklin. "Non-life Insurance Reserve Prediction Using LightGBM Classification and Regression Models Ensemble." Cyber-Physical Systems: Intelligent Models and Algorithms. Cham: Springer International Publishing, 2022. 181-188.

10. Lally, Nathan R., and Brian M. Hartman. "Predictive modeling in long-term care insurance." North American Actuarial Journal 20.2 (2016): 160-183.

11. Grize, Yves-Laurent, Wolfram Fischer, and Christian Lützelschwab. "Machine learning applications in nonlife insurance." Applied Stochastic Models in Business and Industry 36.4 (2020): 523-53