**MTH 522-03: Advanced Mathematical Stats**

**Final Project: Health Insurance Response Prediction**

|  |  |  |
| --- | --- | --- |
| **Team Members** | | |
| Name | ID number | Mail ID |
| Rugwesh Reddy Gankidi | 2195223 | [rgankidi@umassd.edu](mailto:rgankidi@umassd.edu) |
| Gnaneshwar Kethavath | 2192035 | [gkethavath@umassd.edu](mailto:gkethavath@umassd.edu) |
| China Subba Rao Koduri | 2190693 | [ckoduri@umassd.edu](mailto:ckoduri@umassd.edu) |

**Abstract**

Building a model to predict whether the person will be interested in their proposed Health plan/policy given the information about Demographics (city, age, region etc.), Information regarding holding policies of the customer, Recom- mended Policy Information etc. we need supervised models in health insurance process because of the Rapidly growing data volumes-smartphones, smart TVs, and fitness trackers Significant automation capabilities-forecast the types of insurance, and coverage plans, faster and Better Risk Detection-claims processing to policy termination. we utilized decision tree model to predict customer interest in health insurance plans. Evaluation metrics such as accuracy, sensitivity, and specificity were utilized to assess model performance. Findings revealed consistent model accuracy of approximately 77%, with decision trees.

**I. INTRODUCTION**

The project aimed to apply data mining techniques to a real-world dataset to develop a classification model for predicting customer interest in a health insurance plan. Various classification algorithms, such as logistic regression, decision trees, random forests, K Nearest Neighbours (KNN), and Naive Bayes, were explored to find the most suitable model. The Motivation and goal were to leverage demographic information, customer policy details, and recommended policy data to build an effective predictive model for addressing practical applications in the health insurance vertical. Applications: For insurance companies to manage their data and analytics, they must implement emerging technologies like ML and data mining models. It enables insurance businesses to improve operational efficiency while reducing the existing risk of insurance fraud. Some Applications include Price optimization, customer lifetime value, claims processing, risk assessment etc.

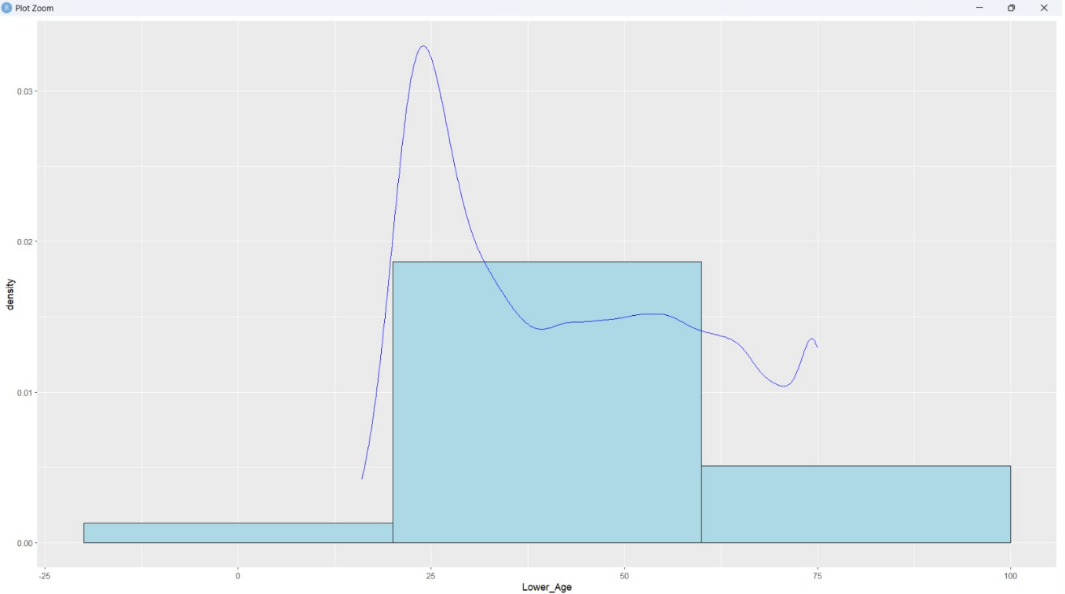
**II. LITERATURE REVIEW & METHODS**

***A. Dataset Information***

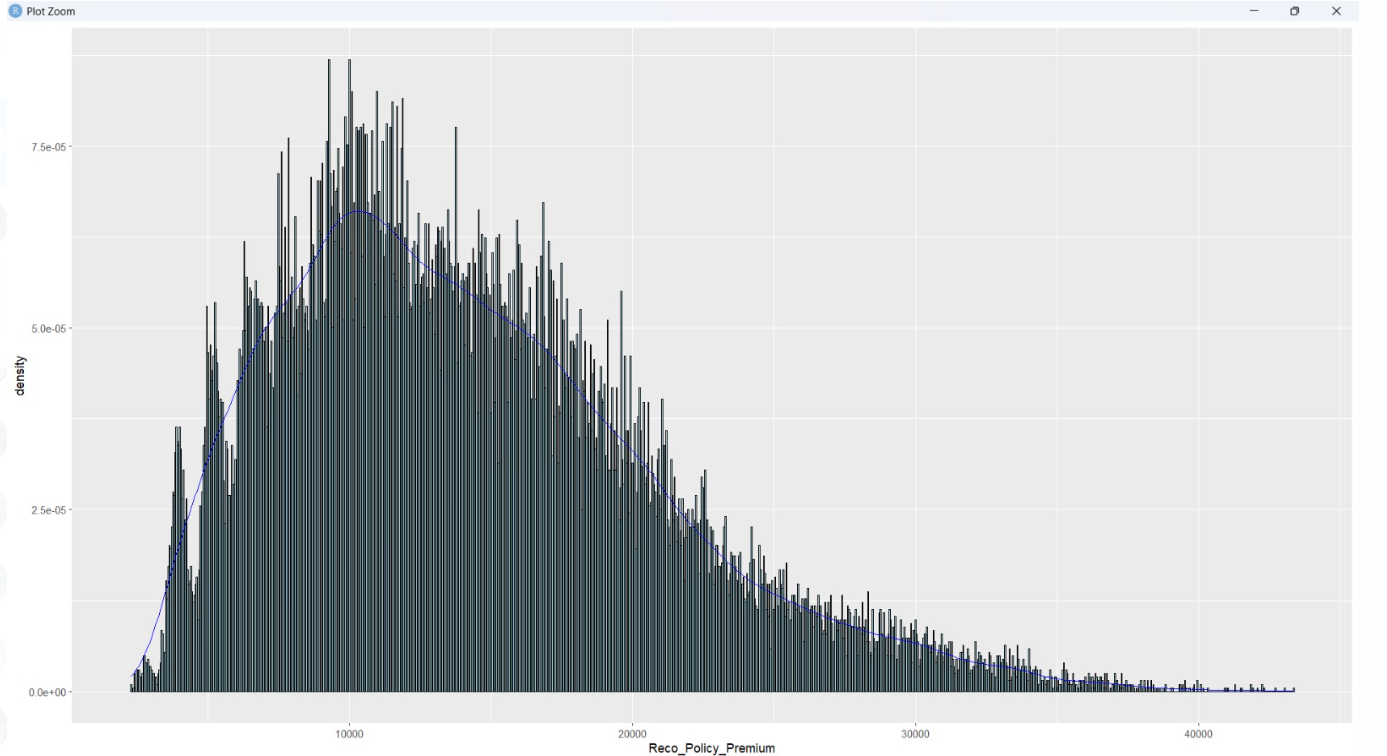
Dataset used in this project is obtained from the kaggle website The dimensions of this dataset are 72,687 rows and 14 columns Features of Dataset: ID: An identifier for each individual/customer. City Code: Code representing the city of residence for the individual. Region Code: Code representing the region of residence for the individual. Accommodation Type: Specifies whether the individual resides in a rented or owned accommodation. Reco Insurance Type: Indicates the type of recommended insurance, whether for an individual or a joint policy. UpperAge: The upper age limit of the individual. LowerAge: The lower age limit of the individual. Is Spouse: Indicates whether the individual has a spouse (Yes/No). Health Indicator: Indicates the health condition of the individual, represented by codes such as X1, X2, etc Holding Policy Duration: Duration of holding the insurance policy, represented in years or categories like ”14+.” Holding PolicyType: the type of the different policies holding. Reco Policy Cat: the different categories of the policy Reco Policy Premium: the premium price of the health insurance Response: Binary variable indicating whether the individual showed interest in purchasing the recommended insurance policy (0 for no, 1 for yes).

***B. Data Visualization***

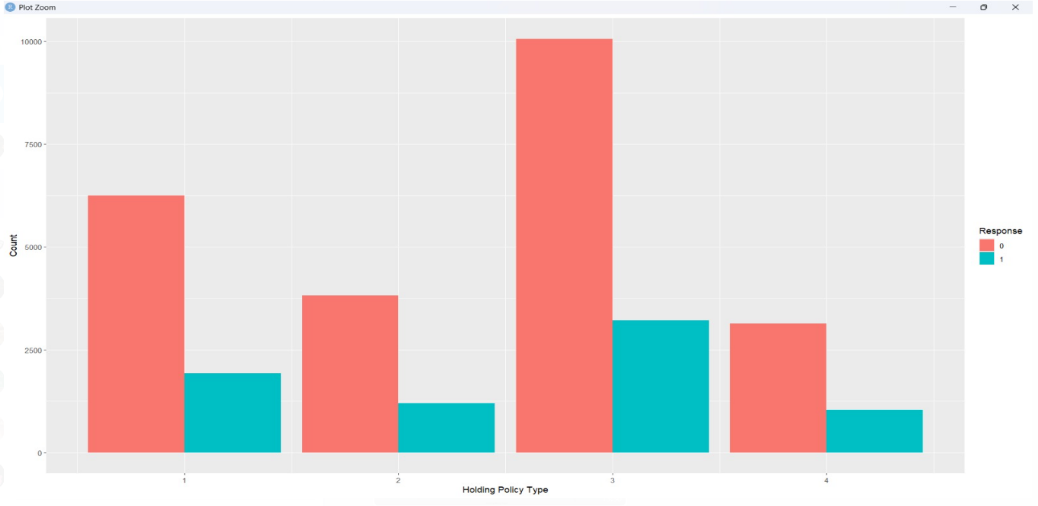
Data visualization is a crucial component of exploratory data analysis as it enables us to uncover patterns and relationships within dataset more effectively. By using visual representations such as histograms, scatter plots, line charts, and Box Plots, analysts can gain insights into the distribution, trends, and dependencies among variables. These visualizations help identify outliers, understand data structures, and detect potential correlations that may guide further analysis. Ultimately, data visualization in EDA enhances data comprehension, aids in hypothesis generation, and supports informed decision-making processes. Below are some of the visual plots for the numerical features in the dataset.



**Fig.no-1:** Age distribution where the majority of the customers are between 25 to 55 years(densely populated) old



**Fig.no-2:** density plot for recommended policy premiums reveals a range between 0 to 40,000, with a mean value of 14,000.



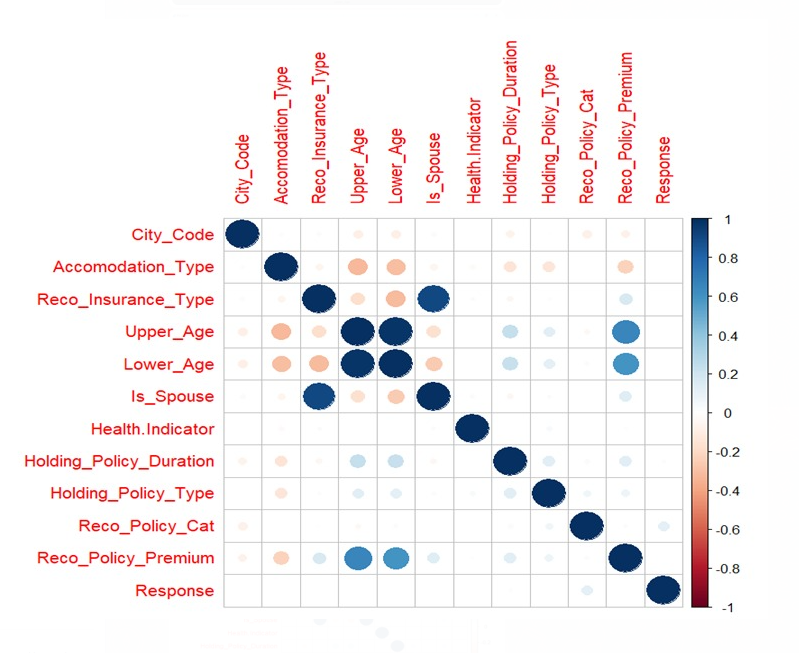
**Fig. No-3:** The bar graph suggests that Policy 3 has the highest count for both output variables 0 and 1, indicating greater interest in the recommended policy plan compared to other types 1, 2, and 4.

***C. Data Preprocessing***

With the initial overview of the data, we encountered the dataset is characterized by significant noise and lack of uniformity, presenting challenges for model development. The initial steps involved addressing missing values through imputation techniques such as imputing with mean to ensure data completeness and reliability. Simultaneously, data cleaning procedures were applied to standardize inconsistent formats. for instance, converting age ranges such as ”14+” to a consistent format like ”15” for better clarity and consistency in the dataset. Further, categorical variables were transformed into a numeric format using label encoding to enable statistical analysis and model training. This transformation enhanced the interpretability and compatibility of the data for data mining algorithms. Additionally, numerical variables underwent scaling to normalize their distributions, mitigating the impact of different measurement scales on model performance. Finally, outlier detection and removal techniques were ap plied to improve the robustness and accuracy of subsequent analyses. By systematically addressing these preprocessing challenges, the dataset was refined into a cleaner, more structured form, laying a solid foundation for subsequent modeling and analysis tasks. These steps exemplify the importance of meticulous data preprocessing in ensuring the quality and reliability of analytical outcomes.

***D. Data Modelling***

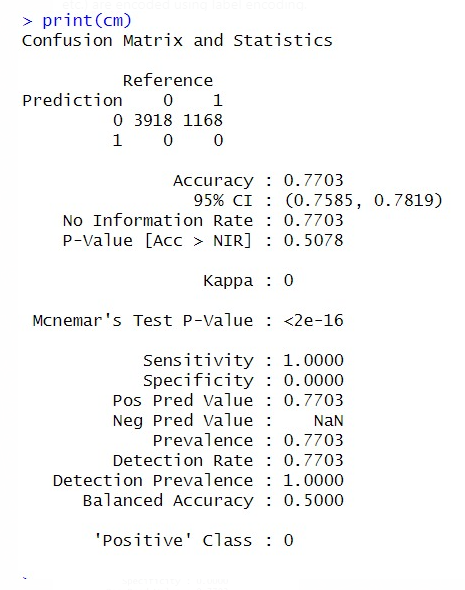
The Decision Tree model was chosen as one of the primary classifiers due to its interpretability and ability to capture complex relationships within the data. The model was trained using the **rpart** function in R, with the target variable being Response, which indicates whether a customer showed interest in the proposed health insurance policy. The remaining features, such as demographic information, health indicators, and policy details, were used as predictors. The cp (complexity parameter) was set to 0.01 to control the growth of the tree and prevent overfitting. This ensured that the tree would stop growing when no further improvements could be made on the training data. After training the Decision Tree model, predictions were generated on the test dataset to assess its performance. A confusion matrix was used to evaluate the model, helping to determine important metrics such as accuracy, sensitivity, and specificity. The confusion matrix provides insight into the true positive, false positive, true negative, and false negative values, which are essential for measuring the performance of classification models.

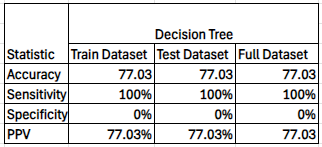


**Fig. No-4:** Based on the figure, it appears there is a positive relationship between age and the Reco Policy Premium column.

**EVALUATION METRICS**

**Accuracy:** Consistent at 77.03% across splits. Sensitivity: 100%, indicating all interested customers are identified correctly. Specificity: 0%, showing all non-interested customers are incorrectly predicted as interested. this also states that , there is a class imbalance problem in the target variableresponse, the model is more biased towards the positive values.



.

**CONCLUSION**

We have started this project by first performing data prepro- cessing step as we found that the dataset is uncleaned, has lot of missing values and null values. Target variable(Response) is unbalanced has more 1’s than 0’s. Lower Age of Customers is densely populated between 18 and 28 years The Premium price is mostly in the range of 10k to 15k Age, Insurance Type, Policy Premium etc. these variables are positively correlated to the target variable Response, compared with the remaining features in the data. After training the models, we found that the decision trees and the naive bayes models performed better in terms of accuracy, followed by logistic regression. There is no big difference for the accuracy score for the training , test and full datasets, for the three models. And have an average accuracy of 72%. The AUC Score for the logistic Regression is about 68% Future studies could explore advanced techniques for class balancing and optimize model execution. We are definitely proud of Preprocessing our data, initially we have achieved 52% accuracy for every model, after preprocessing data effectively our model accuracy has improved to 73%. We would recommend the Naive Bayes model because it strikes a good balance between identifying potential customers and managing resource allocation efficiently. With an accuracy of around 72%, it performs comparably to the Decision Tree but offers some specificity (9-20%), which is crucial for minimizing unnecessary engagement with uninterested leads. Additionally, its high sensitivity (approximately 95%) ensures that most interested customers are captured, making it ideal for maximizing outreach in marketing campaigns. This balance makes Naive Bayes a practical choice for targeted strategies that aim to optimize both lead generation and cost- effectiveness.

**References:**

* 1. https://[www.projectpro.io/article/machine-learning-in-insurance/774](http://www.projectpro.io/article/machine-learning-in-insurance/774)
  2. Dataset:https://[www.kaggle.com/datasets/herambgaidhani/analytics-](http://www.kaggle.com/datasets/herambgaidhani/analytics-) vidhya-jobathon
  3. <https://rpubs.com/zubairishaq9/how-to-normalize-data-r-my-data>
  4. https://[www.r-bloggers.com/2021/04/random-forest-in-r/](http://www.r-bloggers.com/2021/04/random-forest-in-r/)
  5. https://[www.projectpro.io/article/machine-learning-in-insurance/774](http://www.projectpro.io/article/machine-learning-in-insurance/774)