Travel Insurance Purchase Forecast

Zahra Fatah

Dec. 2023

Abstract

This study aims to assist companies in determining the most effective method for encouraging travel insurance purchases. Through an analysis of diverse factors such as age, employment, education, income, family size, health conditions, travel history, and more, the goal is to gauge customer interest in travel insurance. Using a dataset comprising 1887 observations across 9 columns of varied information, this research explores several classification algorithms and ensemble methods such as logistic regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (KNN), Decision Tree, Bagging, Gradient Boosting, Random Forest, and Support Vector Machine and Neural Networ. Among these methods, Random Forest stands out, showcasing an 86% accuracy and an 80% F1 score in predicting customer interest in purchasing travel insurance. The recommended model holds promise for insurance companies, offering valuable insights to precisely target individuals and maximize profitability.

1 Introduction

In an ever-evolving travel landscape, the integration of comprehensive insurance packages has become a pivotal offering by tour and travel companies. Presently, a leading tour & travels company has developed a novel travel insurance package inclusive of COVID coverage. To efficiently target potential buyers within their customer base, the company seeks insights gleaned from historical data. This data comprises records from 2019, capturing the performance and sales of the travel insurance package amongst 1987 previous customers. The primary objective is to construct a predictive model capable of discerning customer interest in purchasing the enhanced travel insurance package.

Dataset Overview: The dataset encapsulates various key features of the customers, providing crucial insights into their inclinations towards the travel insurance package. These features include:

- Age: Reflecting the age demographics of the customers.
- Employment Type: The sector in which the customer is employed.
- Graduate Status: Identification of whether the customer is a college graduate.
- Annual Income: The yearly income of the customer in Indian Rupees.
- Family Size: Number of members within the customer's family.
- Presence of Chronic Disease: Noting any major health conditions like diabetes, high blood pressure, asthma, etc.
- Frequent Flyer Status: Derived from the customer's history of booking air tickets on at least four different instances within the last two years (2017-2019).
- Travel History: Indication of whether the customer has traveled to a foreign country.
- Travel Insurance: Identification of customers who purchased the travel insurance package during the introductory offering in 2019.

By analyzing these distinct features, the company wants us to create a robust predictive model that accurately anticipates customer interest in acquiring the travel insurance package with COVID coverage. This model should facilitate targeted marketing strategies, enabling the company to tailor its offerings to customers most likely to embrace the new insurance package, thereby optimizing sales and customer satisfaction.

2 Exploratory Data Analysis

The dataset comprises 1987 customer observations, with 100 observations allocated for testing the best-fitted model. Among these, 80% (1887 observations) are assigned to the training set, while 378 observations are designated for validation tests. As delineated in the introduction, the dataset contains 9 features, with the 'TravelInsurance' feature serving as the target variable, aimed at predicting whether new customers will purchase the travel insurance. Among the features, 'Age,' 'AnnualIncome,' and 'FamilyNumber' are numerical variables, while the remaining six features are categorical.

Our initial exploration involves examining both univariate and bivariate aspects of the data. This process aims to familiarize ourselves with the dataset, paving the way to employ various models for fitting the data and determining the most suitable one. Through this exploratory phase, we seek to gain insights that will facilitate the development of a robust predictive model for anticipating customer interest in the travel insurance package with COVID coverage.

2.1 Univariate EDA

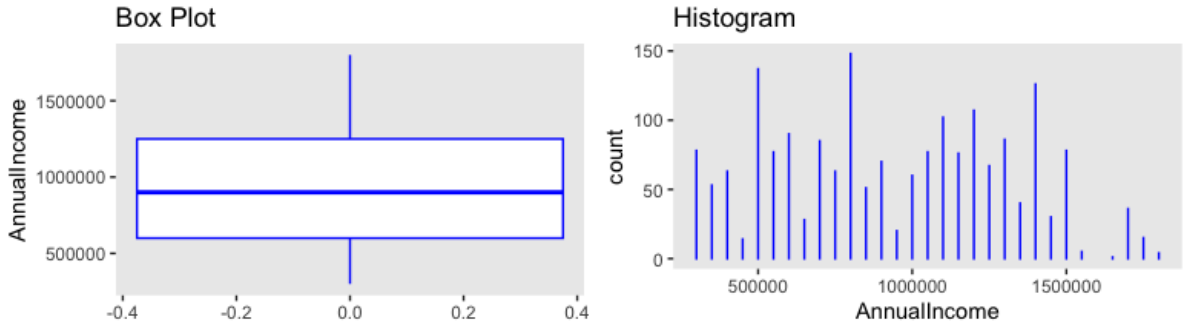

Figure 1: Annual Income Boxplot & Histogram

Minimum and maximum annual income of the customer is 300,000 and 1,800,000 respectively and average salary is around 900,000. The number of customer with high incom is low.

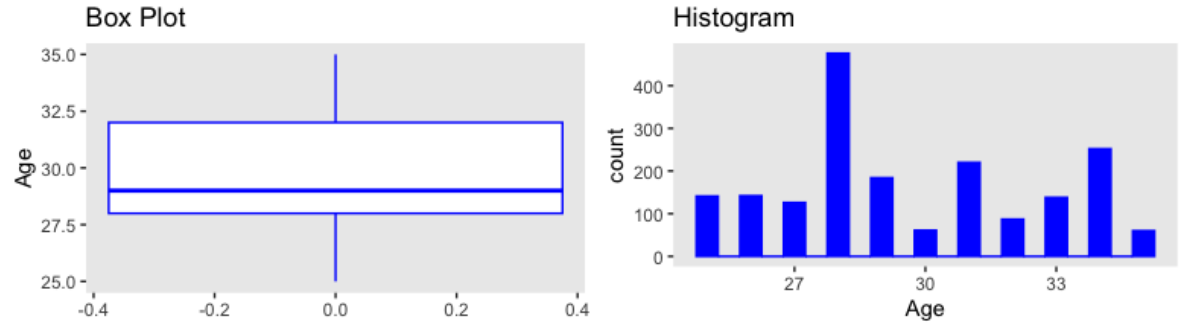

Figure 2: Age Income Boxplot & Histogram

Minimum and maximum age of the customer is 25 and 35 respectively and average age is 29.64. The number of customers who are 28 years old is high.

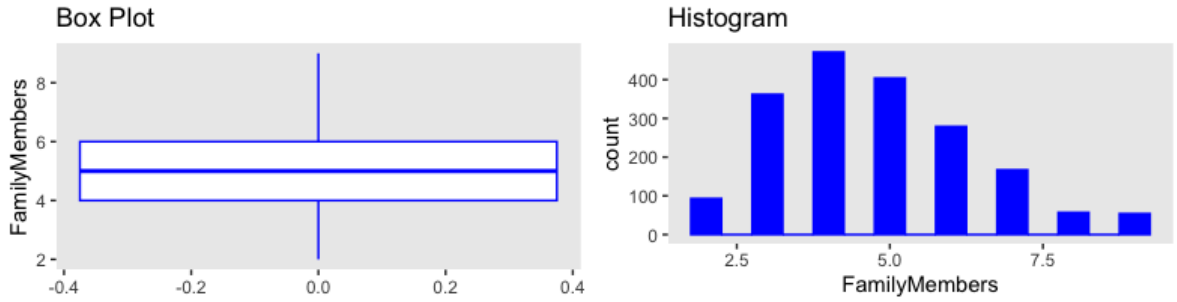

Figure 3: Family Members Boxplot & Histogram

Minimum and maximum number of family members in the customer's family is 2 and 9 respectively with average of around 5. Most of the customers have in general more than 4 family members in the family. The distribution of family members is close to normal.

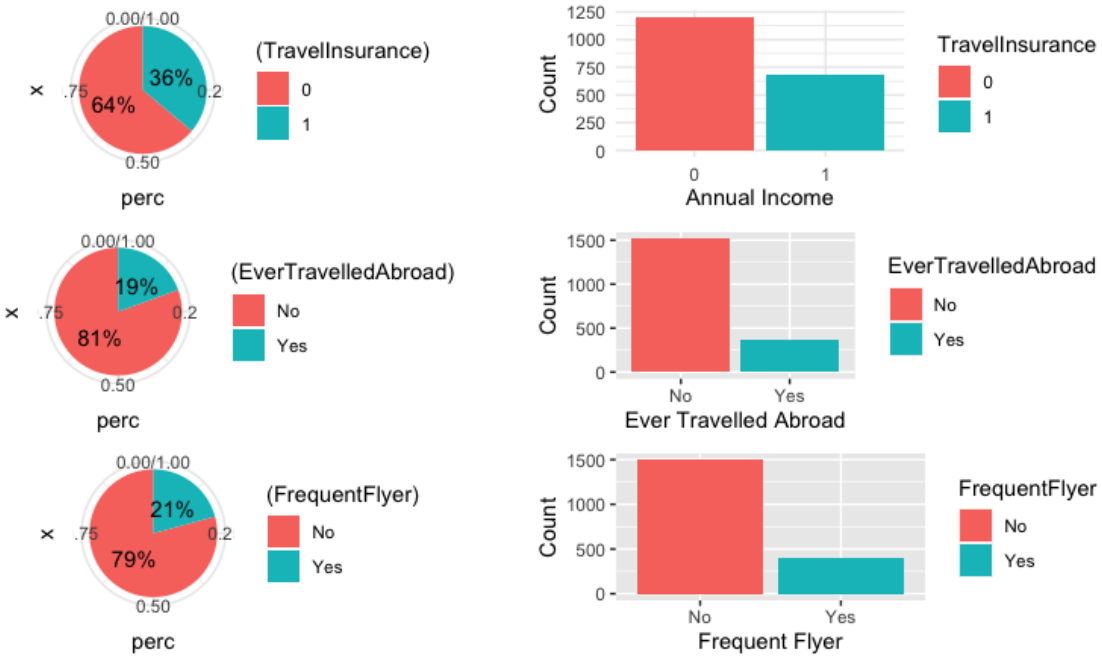

Figure 4: Travel Insurance, Ever Travel Abroad, and Frequent Flyer pie charts and bar plots

36% of customers bought travel insurance. Only 19% of the all customers have an abroad travel experience. Most of the customers are not frequent flyers, only 21% of the all customers are frequent flyers.

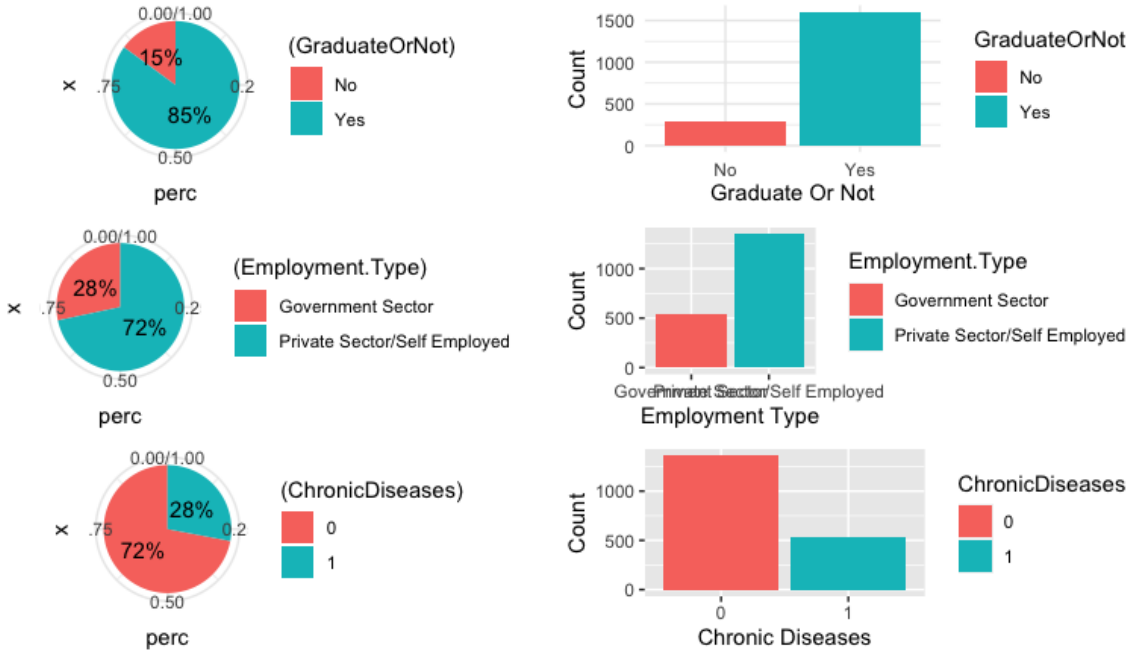

Figure 5: Graduate or Not, Employment Type, Chronic Diseases pie charts and bar plots

85% of customers hold college degrees. The count of customers employed in the governmental sector is notably lower compared to those in the private sector or self-employment, constituting 72% of the customer base. Out of total customers only 528 customers are suffering from major disease but 1359 of the customers were not suffering from any major diseases.

2.2 Bivariate EDA

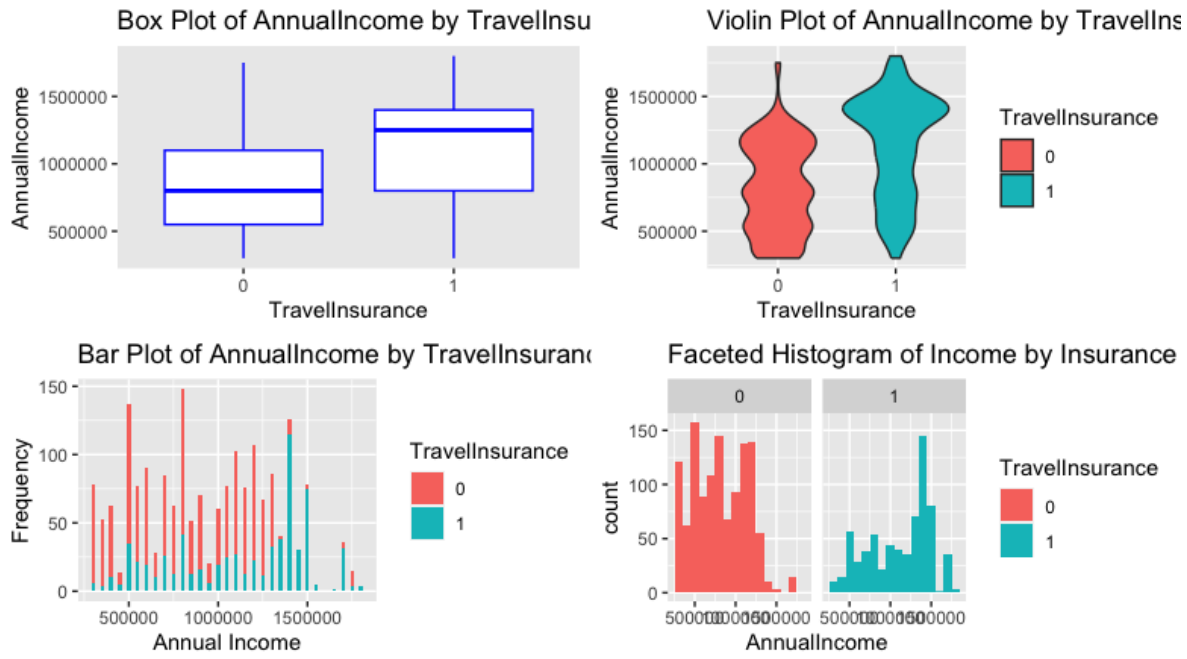

Figure 6: Annual Income & Travel Insurance

The median income for people buying travel insurance surpasses that of those who don't. Customers with high salaries are almost certain to buy travel insurance, while those with lower salaries are almost certain not to purchase it.

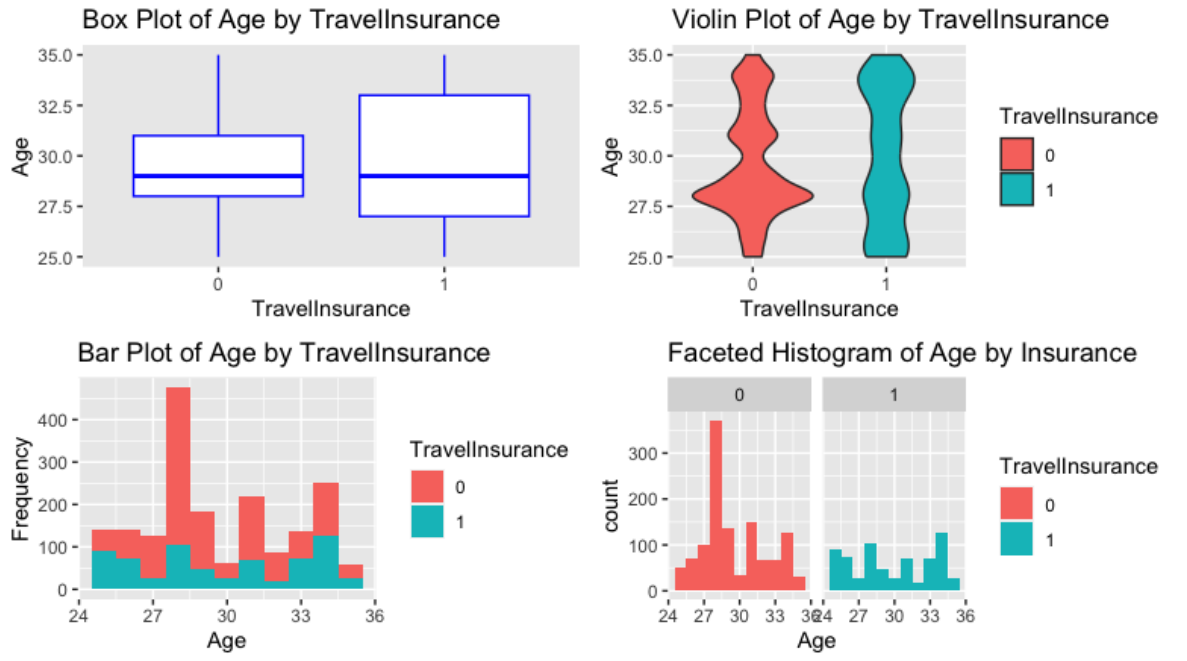

Figure 7: Age & Travel Insurance

A customer aged 34 has the highest number of purchased travel insurance policies. The number of customers who did not purchase insurance is significantly higher at the age of 28 compared to those who bought it. Generally, customers at a lower or higher age are more likely to purchase travel insurance

than those in the middle age bracket. The distribution of age in both classes does not resemble a normal distribution.

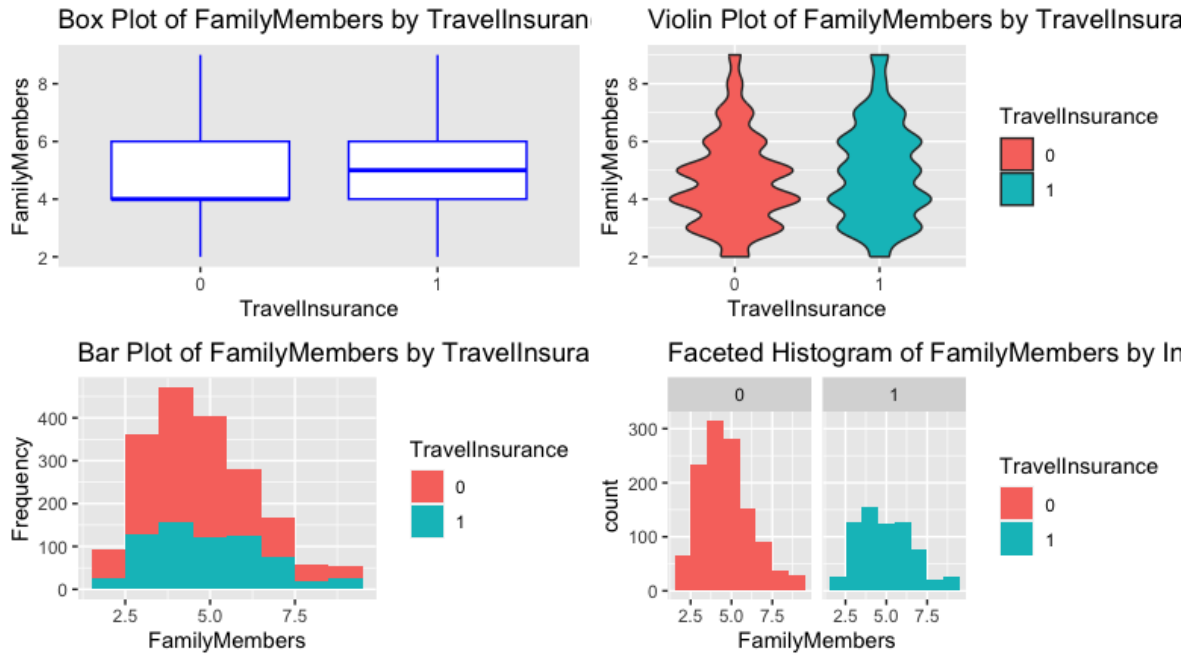

Figure 8: Family Members & Travel Insurance

The median number of family members in the family for customers who purchase travel insurance is more than those who did not purchase the travel insurance in 2019. Customers with number of family member 4 are the one that has the highest number of travel insurance purchased. The distribution of age in both classes are close to normal distribution.

2.3 Model Performance Metrics

Statistical metrics play pivotal roles in evaluating the performance of classification models. Here are some key metrics used to validate our fitted models:

Accuracy: This metric measures the overall correctness of a model. It calculates the ratio of correctly predicted instances to the total instances in the dataset.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Precision: Precision assesses the accuracy of positive predictions made by the model. It represents the ratio of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall (Sensitivity): Recall, also known as sensitivity, measures the model's ability to correctly identify true positives. It's the ratio of correctly predicted positive observations to the actual positives in the dataset.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall, especially when there is an uneven class distribution.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Specificity: Specificity measures the model's ability to correctly identify true negatives. It's the ratio of correctly predicted negative observations to the actual negatives in the dataset.

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

While accuracy measures the overall correctness, precision and recall focus on specific aspects of the model's performance related to positive predictions and true positive identification, respectively. The F1 score balances precision and recall, and specificity evaluates the ability to correctly identify true negatives.

Using several methods and model fitting on data provides a robust assessment of the data and ensures that the chosen model is well-suited for the underlying patterns and characteristics present in the dataset. This approach helps in making a more informed decision about the most appropriate model for the specific problem. We are trying Logistic regression, LDA, QDA, DT, RF, XGBoost, KNN, Neural Network in this paper.

3 Logestic Regression

3.1 Model Specification and fitting

Since we have binary classification problem and the sample size is large, logistic regression is considered one of the best methods for fitting our model in such cases. Upon fitting the model, it becomes evident that the variables ChronicDiseases, GraduateOrNotYes, and Employment.TypePrivate Sector/Self Employed are not statistically significant. Their p-values exceed the level of significance (0.05). Utilizing a backward model approach, we systematically remove each of these variables starting with the one having the largest p-value. This process aims to refine the model by eliminating variables that do not significantly contribute to predicting the purchase of travel insurance.

3.2 Model validation

In Figure 9 and Figure 10, the results of the statistical metrics support this observation, as there are no differences in the outcomes when these variables are removed.

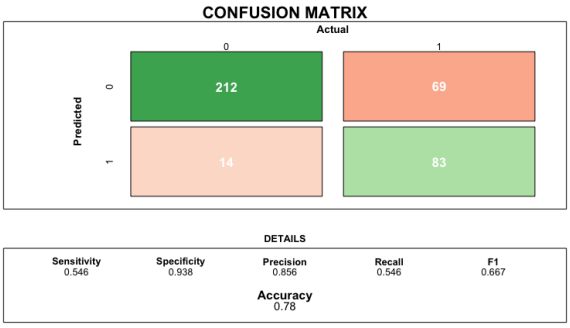

Figure 9: Logistic regression Confusion Matrix

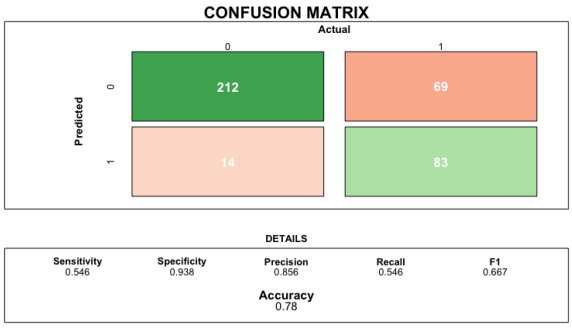

Figure 10: Logistic regression with feature selection

4 Linear & Quadratic Discrimination Analyses

LDA seeks to find the linear combination of features that best separates different classes in the dataset. It assumes that the classes have a similar covariance matrix and that the data is normally distributed within each class. LDA projects the data onto a lower-dimensional space, aiming to maximize the between-class variance while minimizing the within-class variance. It's especially useful when the classes are well-separated and the assumptions of normality and equal covariance hold. With a larger sample size of 1887 observations, it might be reasonable to assume approximate normality within each class for the purposes of applying methods like LDA. QDA is an extension of LDA that relaxes the assumption of equal covariance matrices for each class.

4.1 Model Specification and fitting

We applied LDA and QDA model on our data for all 8 features.

4.2 Model validation

The results of the statistic metrics for the LDA and QDA methods are depicted in Figures 11 and 12, respectively. We observe that, based on accuracy scores, the performance of LDA is nearly identical to

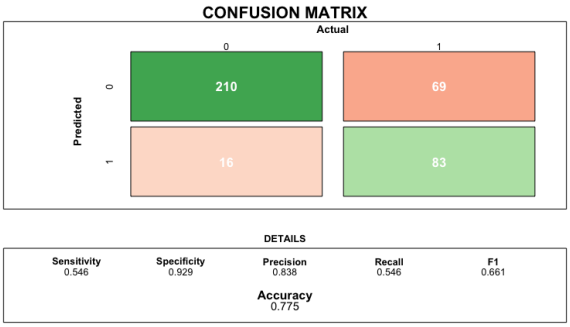

Figure 11: LDA Confusion Matrix

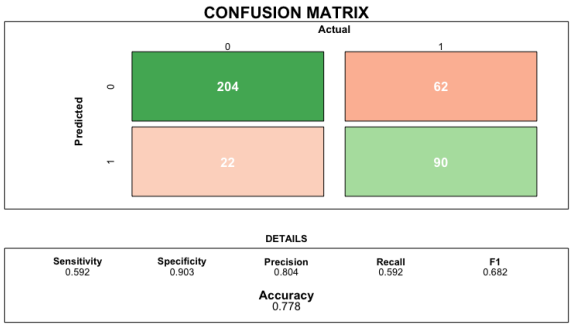

Figure 12: QDA Confusion Matrix

that of QDA, but LDA demonstrates higher precision and lower sensitivity compared to QDA.

5 KNN

K-Nearest Neighbors (KNN) presents a straightforward and versatile approach in machine learning, suitable for both classification and regression tasks. Its simplicity in implementation and non-parametric nature, accommodating various data types without making assumptions about data distributions, makes it an attractive choice. KNN's resilience to outliers, interpretability in decision-making, and lack of a training period further enhance its value. However, considerations such as the choice of the optimal 'k' value and computational expenses during testing phases should be noted.

5.1 Model Specification and fitting

Using tuning and cross validation, accuracy was used to select the optimal model using the largest value. The final value used for the model was k = 17.

5.2 Model validation

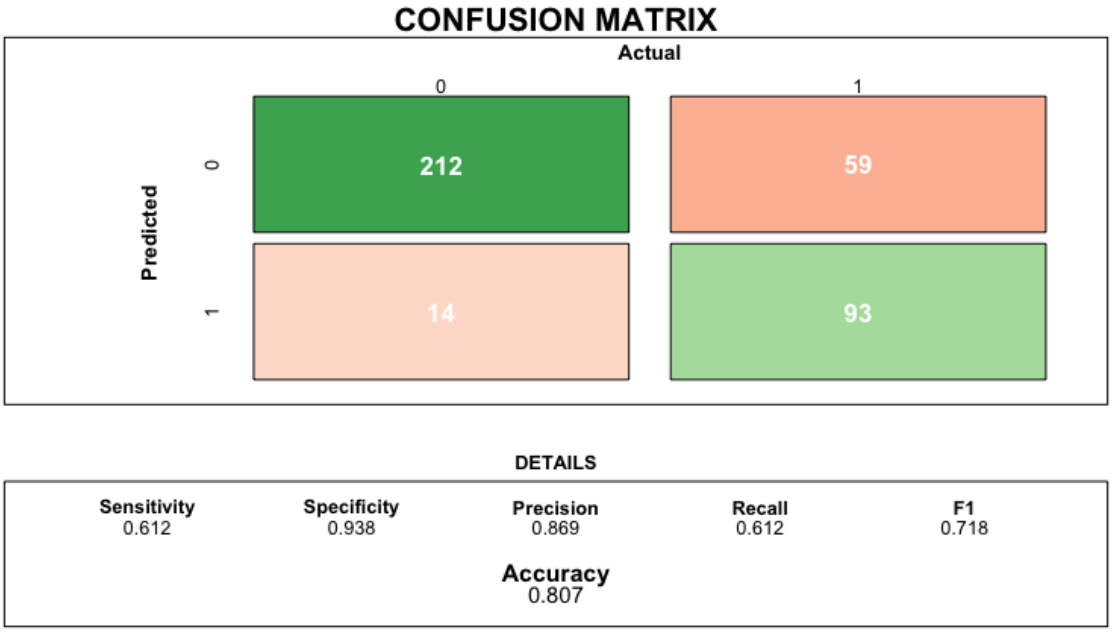

Figure 13: KNN Confusion Matrix

According to the confusion matrix results, the KNN model exhibits higher accuracy compared to logistic regression, LDA, and QDA.

6 Decision Tree & Pruned DT

6.1 DT

Decision Trees offer several advantages that make them valuable in various machine learning scenarios. One key advantage is their inherent interpretability and ease of understanding. DT models mimic human decision-making processes, generating simple, visual representations that are intuitive to comprehend. They can handle both numerical and categorical data, require minimal data preparation (like normalization or scaling), and are robust to outliers. Additionally, Decision Trees implicitly perform feature selection by identifying the most relevant features for classification. They can handle non-linear relationships between features and the target variable without requiring complex transformations. Their ability to handle interactions and automatically select important features makes them particularly useful in exploratory analysis and providing insights into the data. These factors contribute to their popularity and utility in diverse domains, especially when transparency and interpretability are essential.

6.1.1 Model Specification and fitting

The Insurance Decision Tree comprises 154 terminal nodes, With a residual mean deviance of 0.763, it demonstrates a moderate alignment with the dataset; also, its misclassification error rate stands at 0.164. While the extensive structure of 154 terminal nodes implies a complex decision-making process, the relatively high misclassification rate suggests areas where refinement is necessary to enhance the model's predictive accuracy.

6.1.2 Pruned DT

we use cross validation method on the training set in order to determine the optimal tree size. by checking the deviance error which becomes constant after tree size = 5,we chose the best tree size = 5. we pruned our dt by best tree size. The residual mean deviance is 0.901.The residual mean deviance is 0.172, we can see that both Residual mean deviance and Misclassification error rate are greater for best tree size = 5.

6.2 Model validation

comparing Dt and pruned dt statistical metrics, we can see an improvement in our model fitting by prunning the tree.

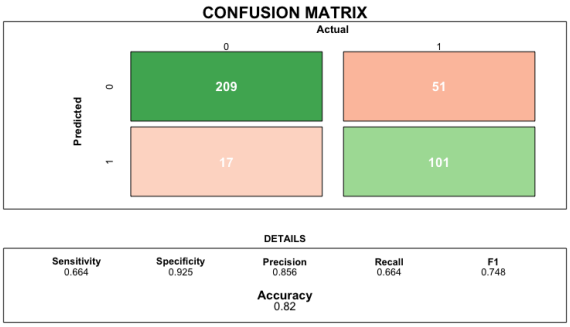

Figure 14: Decision Tree Confusion Matrix

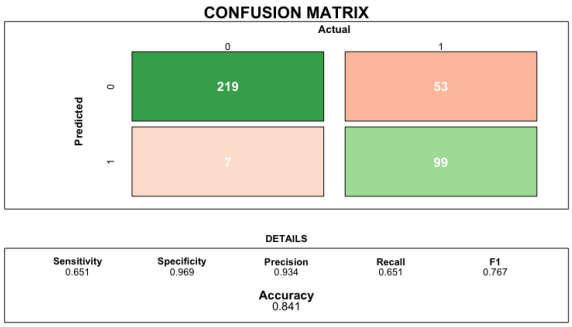

Figure 15: Pruned DT Confusion Matrix

7 Bagging & Random Forest

RF builds multiple decision trees during training and merges their predictions to improve accuracy and reduce overfitting. Each tree in the forest operates independently, making predictions. The final predic-

tion is often the average (regression) or majority vote (classification) of all trees. During the training, randomness is injected during tree construction, using random subsets of features and bootstrapped samples of the dataset. The advantages of RF are: robust to overfitting, effective with large datasets, handles missing values well, and provides feature importance. RF can be slow with extensive data and may not perform as well with noisy data.

7.1 Model Specification and fitting

Due to the high variance observed in the Decision Tree (DT) method, we've implemented the Bagging method to address this issue. The presence of out-of-bag (OOB) data eliminates the necessity for cross-validation. Additionally, we've leveraged the Bagging method to enhance interpretability by identifying important variables within our model. With an 'mtry' value of 8, the OOB error rate stands at 21.4

Furthermore, to mitigate the correlation problem, we've utilized Random Forest (RF) with 'mtry' set to $mtry = \text{round}(\sqrt{8}, 0) = 3$, which results in 'mtry = 3.' The OOB error rate achieved with RF is 17.76

AnnualIncome, Age, FamilyMembers, Employment.Type, EverTravelledAbroadYes are important variables in Bagging and RF. SO annual income is the key factor that will decide weather the person will buy the travel insurance or not.

7.2 Model validation

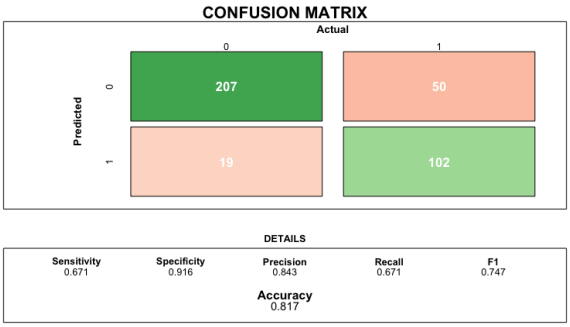

Figure 16: Bagging Confusion Matrix

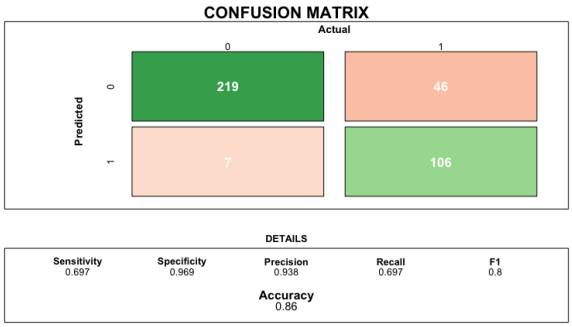

Figure 17: Random Forest Confusion Matrix

As anticipated, the accuracy of Random Forest (RF) surpasses both the accuracy of the Pruned Decision Tree and the unpruned DT.

8 XGBoost

XGBoost is a gradient boosting algorithm that sequentially builds trees, focusing on correcting errors made by previous models. It combines weak learners (usually decision trees) to form a strong learner by boosting their performance iteratively. It Utilizes a gradient descent algorithm to minimize the loss function, adding trees that minimize the residual errors. Exceptional performance, efficient computation, handles missing data, prevents overfitting via regularization, and provides feature importance are some advateges of XGBoost Method.

8.1 Model Specification and fitting

We train our model with all features using cross validation.

8.2 Model validation

We can see the statistical metrics for XGBoost model in Figure 18.

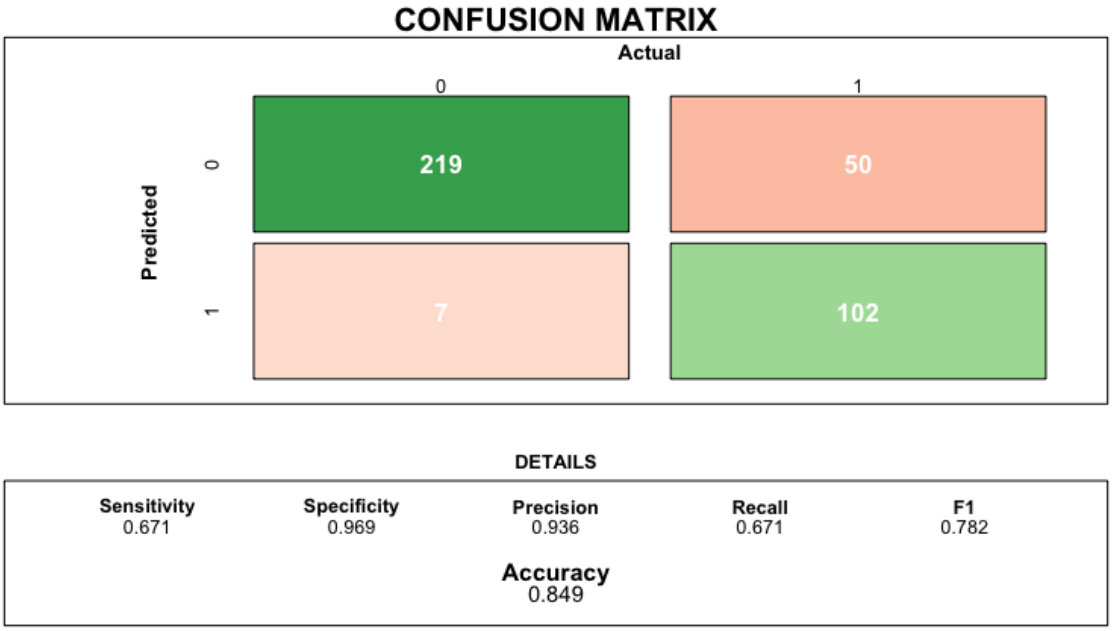

Figure 18: XGBoost Confusion Matrix

9 SVM

Support Vector Machine (SVM) offers several advantages in machine learning. It's highly effective in high-dimensional spaces, making it suitable for scenarios with numerous features. SVM is proficient in handling both linear and non-linear data through the use of different kernel functions, allowing it to model complex relationships effectively. Moreover, SVM is less prone to overfitting due to its margin maximization principle, which aims to find the best decision boundary while maintaining a considerable gap between classes. Additionally, SVM's versatility in handling various data distributions and its ability to work well with small to medium-sized datasets contribute to its usefulness across diverse problem domains.

9.1 Model Specification and fitting

9.1.1 Linear Kernel

We performed model tuning on the SVM using 10-fold cross-validation to identify the optimal cost parameter. The best parameter found was a cost value of 0.01, resulting in the best performance error of 0.2366. The error rate in the training data is 0.24045, and in the test data is 0.2169.

9.1.2 Radial

Kernel We performed model tuning on the SVM using 10-fold cross-validation to identify the optimal cost and gamma parameters. The best parameter found was a cost value of 10 and gamma value of 0.1, resulting in the best performance error of 0.1829.

The training error for the radial kernel (0.1776) is lower than that of the linear kernel (0.24045). Also, the test error for the radial kernel (0.1746) is lower than that of the linear kernel (0.2169). Consequently, these findings suggest that the radial kernel is more effective for our insurance data.

9.1.3 Polynomial Kernel

We performed model tuning on the SVM using 10-fold cross-validation to identify the optimal cost and degree parameters. The best parameter found was a cost value of 5 and degree value of 3, resulting in the best performance error of 0.1856.

9.2 Model validation

Upon comparing the training and test errors across linear, radial, and polynomial kernels, it becomes evident that the radial kernel exhibits the best performance. We can see this result by checking the accuracy of each model.

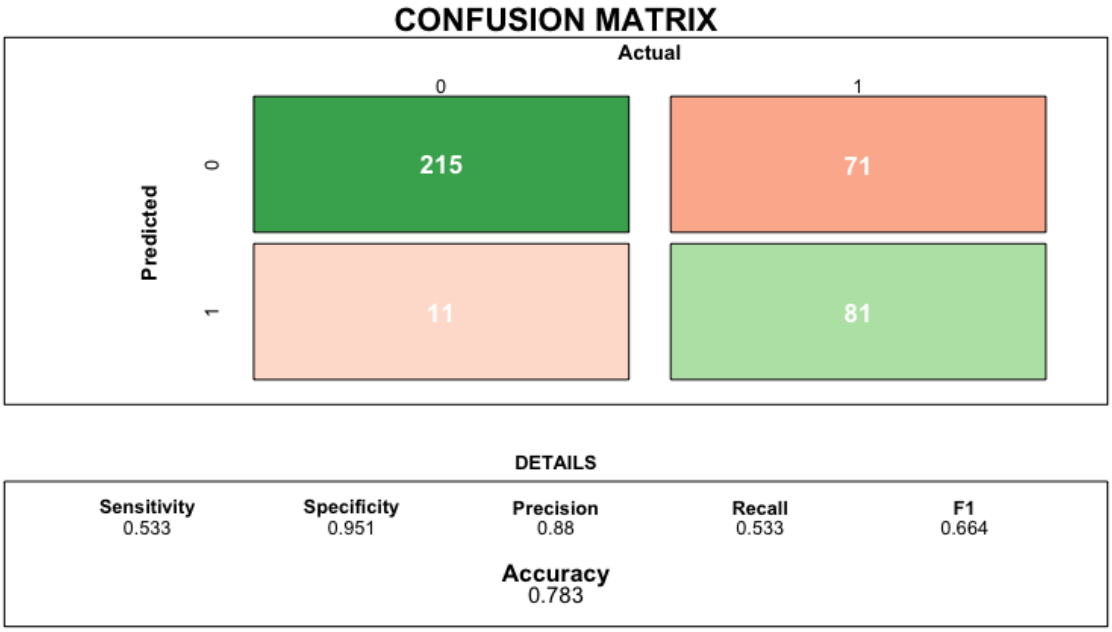

Figure 19: SVM with Linear Kernel

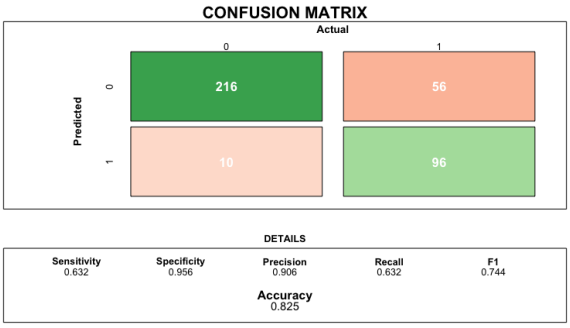

Figure 20: SVM with Radial kernel

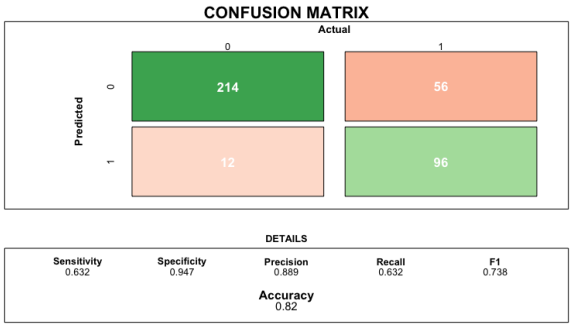

Figure 21: SVM with polynomial kernel

10 Neural Network

Neural networks are able to learn from complex, non-linear relationships within data makes them exceptionally powerful for tasks involving pattern recognition, image and speech recognition, and natural language processing. Neural networks excel in handling large volumes of data, and their adaptability to various problem domains makes them versatile. Additionally, their capacity for parallel processing and distributed computation enables faster training on specialized hardware like GPUs, leading to enhanced performance in complex tasks. Despite their complexity and demand for computational resources, neural networks remain at the forefront of cutting-edge advancements in AI and deep learning.

10.1 Model Specification and fitting

We performed model tuning on the NN using cross-validation to identify the optimal decay and size parameters. The best parameter found was a size value of 6 and decay value of 0.1, resulting in the best accuracy of 0.7840.

10.2 Model validation

We can see the statistical metrics for NN model in Figure 22.

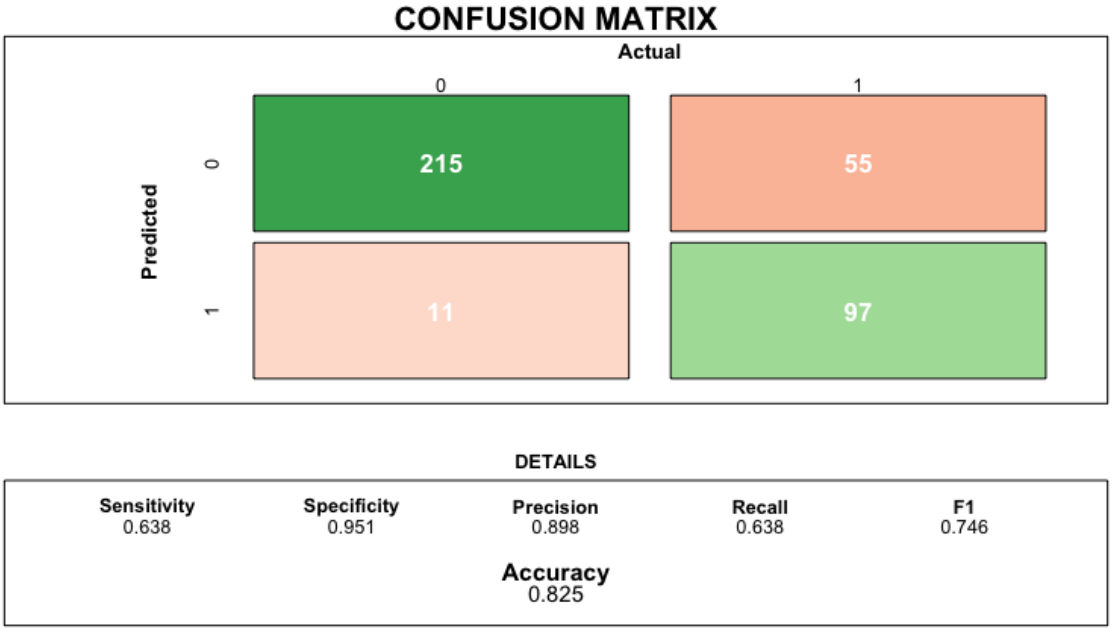

Figure 22: Neural Network Model

11 Conclusion

Based on the metrics obtained from various models, it's evident that each model demonstrates its strengths in different aspects of classification. The Random Forest model displays the highest accuracy and F1 score, indicating its robustness in overall performance. Additionally, the Pruned Decision Tree shows commendable precision while maintaining a relatively high accuracy. However, it's essential to consider trade-offs between precision, recall, and specificity based on the context of the problem at hand. For instance, the SVM Radial model showcases a good balance between accuracy and recall, while the Neural Network (NN) exhibits notable precision. Understanding these nuanced differences is crucial in selecting the most appropriate model for deployment, ensuring optimal performance aligned with specific requirements and objectives.

Model	Accuracy	F1	Precision	Recall	Specifity
Logistic Regression	0.78	0.667	0.856	0.546	0.938
Linear Discriminant Analysis	0.775	0.661	0.838	0.546	0.929
Quadratic Discriminant Analysis	0.778	0.682	0.804	0.592	0.903
K-Nearest Neighbors	0.807	0.718	0.869	0.612	0.938
Decision Tree	0.82	0.748	0.856	0.664	0.925
Pruned DT	0.841	0.767	0.934	0.651	0.969
Bagging	0.817	0.747	0.843	0.671	0.916
Random Forest	0.86	0.8	0.938	0.697	0.969
XGBoost	0.849	0.782	0.936	0.671	0.969
SVM Linear	0.783	0.664	0.880	0.533	0.951
SVM Radial	0.825	0.744	0.906	0.632	0.959
SVM Polynomial	0.820	0.738	0.889	0.632	0.947
NN	0.825	0.746	0.898	0.638	0.951

Lastly, we utilized our top-performing model, Random Forest, to predict outcomes for the withheld 100 observations. Subsequently, we saved the resulting labels in a CSV file.

12 suggestion

The tour and travel company can leverage the predictive capabilities of the Random Forest model to identify potential customers interested in purchasing travel insurance. Additionally, focusing on crucial features highlighted by logistic regression, Decision Trees (DT), Random Forest (RF), and XGBoost, such as customers' annual income, age, family size, Employment Type, and travel history, would be beneficial. Incorporating these factors into future strategies could optimize the company's marketing efforts, tailoring offerings to better meet the preferences and needs of potential customers.