TRAVEL INSURANCE PREDICTION

Problem Setting:

Travel insurance is used to cover the costs and losses associated with traveling. It is a useful protection for those traveling domestically or abroad. An intelligent model is built that can predict if the customer will be interested to buy the travel insurance package. The solution offered by the insurance policy for the customers may be used for customer-specific advertising. The company would require knowing which customers will be interested to get an insurance based on the database history. Travel insurance may include several types of coverage. Some of the coverages that the travel insurance includes are trip cancellation or interruption coverage, baggage and personal effects coverage, medical expense coverage, and accidental death or flight accident coverage. Also, some travel insurance policies may duplicate existing coverage from other providers or offer protection for costs that are refundable by other means.

Problem definition:

The goal of this analysis is to identify the best data mining model and predictors for exactly predicting whether the customer has bought the travelers insurance or not. The implication is to conclude the prediction of buying the insurance based on the factors of the customers like annual income, frequent flyers, employment type, salary, etc. How do the factors influence the target? How is it related to the target? Which predictors don't belong to the model? Influences the prediction.

Data Sources:

The travel insurance dataset has been taken from Kaggle Data, an open-source repository for data mining. The link to the data set is https://www.kaggle.com/datasets/tejashvi14/travel-insurance-prediction-data

Data Description:

The dataset has 10 columns out of which 9 are attributes and 1 is a target variable 'TravelInsurance', which depicts whether the customer has bought the travel insurance or not by denoting '1' or '0'. The total number of records is 1986.
The Variable names & Description are as follows:

- Age- The age of the customer
- Employment Type- The sector in which the customer is employed.
- GraduateOrNot- Whether the customer is a college graduate or not.
- AnnualIncome- The yearly income of the customer
- FamilyMembers- Number of members in the customer's family
- ChronicDisease- Whether the customer suffers from any major disease (Diabetes/High BP or Asthma, etc.)
- FrequentFlyer- Derived data based on customer's history of air tickets (At least 4 Different Instances in The Last 2 Years)

- EverTravelledAbroad- Has the customer ever traveled abroad
- TravelInsurance- Did the customer buy travel insurance

Data Understanding:

The original dataset consists of 9 attributes, 1987 records, and the target variable 'Travel insurance'. It was discovered by examining the attribute data types that this dataset contains both numerical and category characteristics. All categorical variables, nevertheless, are binary in nature, meaning that each category value may only be either "0" or "1". The binary nature of the target variable "TravelInsurance" means that a value of "0" indicates that the customer did not get travel insurance, while a value of "1" indicates that the customer got travel insurance. There were 1277 occasions when travelers did not obtain travel insurance and 710 instances where they did.

Data Pre-processing:

The variable "Unnamed: 0," which served as a unique identifier for each entry, was eliminated since it was unneeded. After deleting duplicate records from the data as a result of the aforementioned data cleaning stages, 1249 instances, and 8 characteristics remained to be used in the next data mining steps.

Data Summary Statistics:

Age	Employment Type	GraduateOrNot	AnnualIncome	FamilyMembers	ChronicDiseases	FrequentFlyer	EverTravelledAbroad	TravelInsurance
.000000	1249.000000	1249.000000	1.249000e+03	1249.000000	1249.000000	1249.000000	1249.000000	1249.000000
.755805	0.701361	0.838271	9.345476e+05	4.890312	0.333066	0.236189	0.195356	0.386709
.921039	0.457844	0.368350	3.607293e+05	1.762313	0.471499	0.424910	0.396634	0.487191
.000000	0.000000	0.000000	3.000000e+05	2.000000	0.000000	0.000000	0.000000	0.000000
.000000	0.000000	1.000000	6.000000e+05	4.000000	0.000000	0.000000	0.000000	0.000000
.000000	1.000000	1.000000	9.000000e+05	5.000000	0.000000	0.000000	0.000000	0.000000
.000000	1.000000	1.000000	1.200000e+06	6.000000	1.000000	0.000000	0.000000	1.000000
.000000	1.000000	1.000000	1.800000e+06	9.000000	1.000000	1.000000	1.000000	1.000000

Data exploration:

After identifying the relationships, we developed a heat map of the characteristics that were highly predictive of whether a customer would purchase travel insurance. This heat map shows that a customer's purchase of travel insurance relates to their annual income and ever traveling abroad.

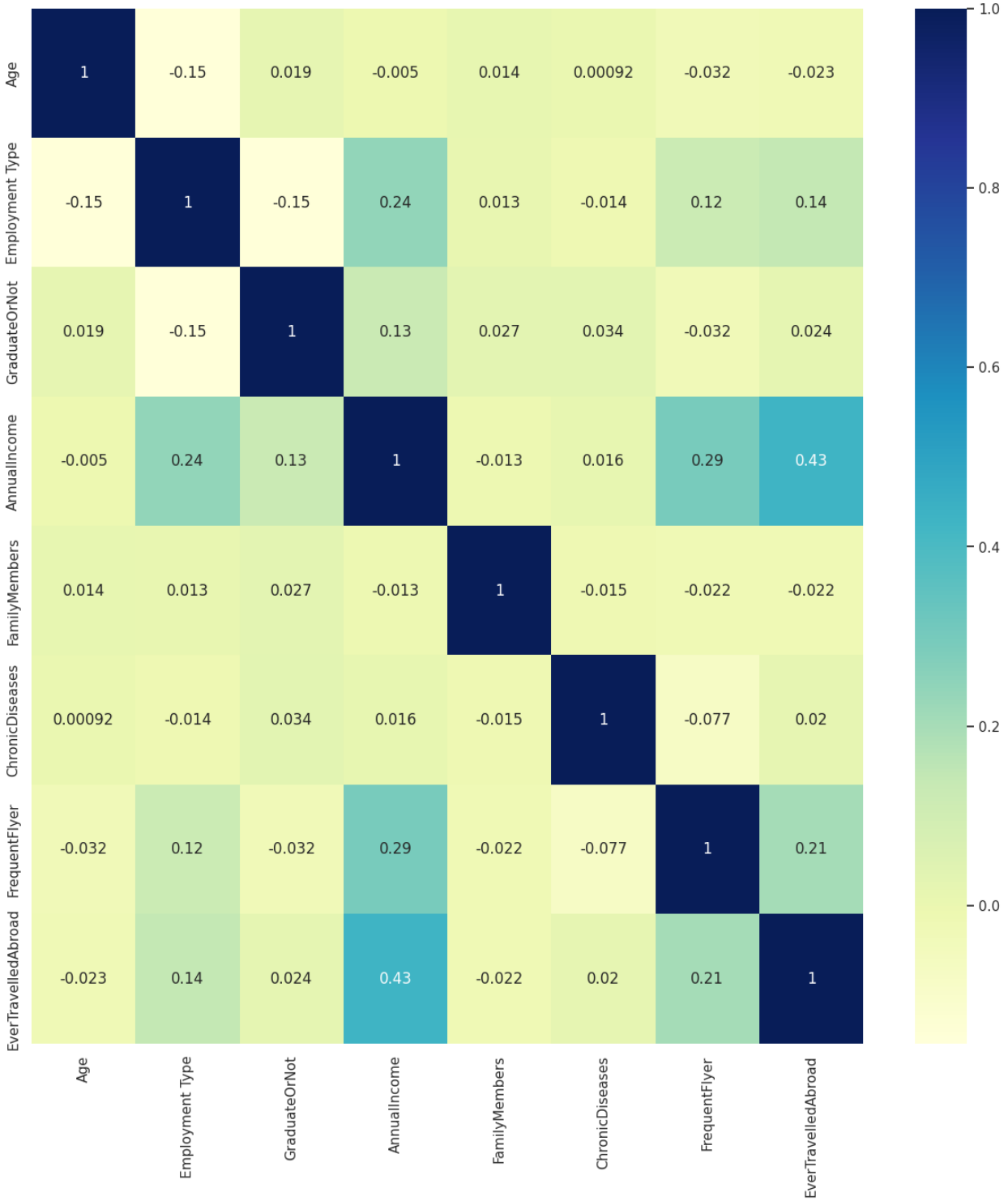

Fig.1 Heatmap on Predictor variables

Also, we employed a pair plot to enhance the visualization of helpful characteristics, such as annual income, ever-traveled abroad, employment type, and frequent flyer.

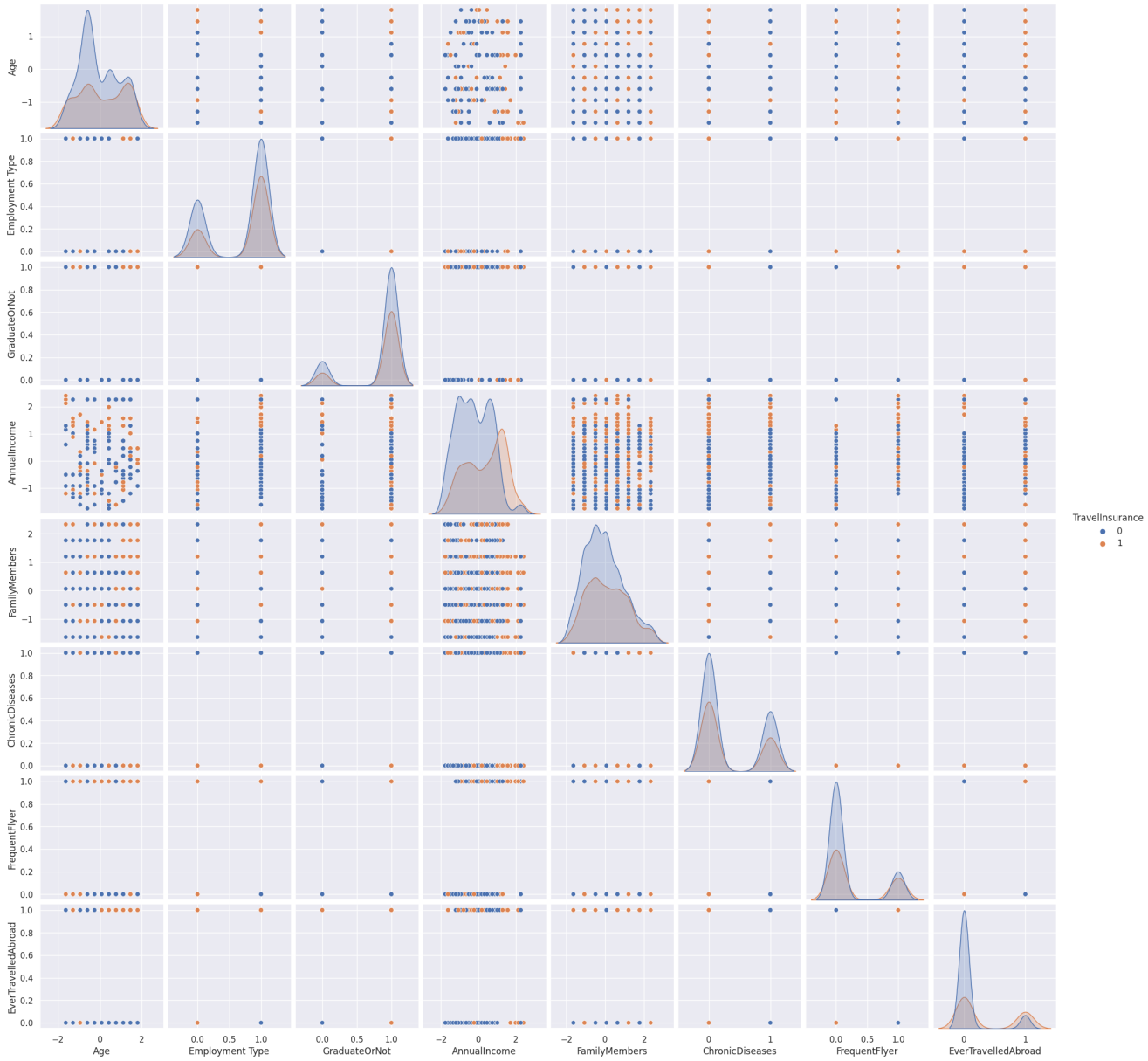

Fig.2 Pair plot on the predictor variables

Using exploratory data analysis, the numerical and categorical variables were looked at independently. The following is a list of the findings from these evaluations:

Categorical Variables:

As the unique values for each of the categorical variables were discovered, it was discovered that all six categorical categories including the target variable 'TravelInsurance' were binary in nature. The histogram plot reveals the count of variables having 0 and 1.

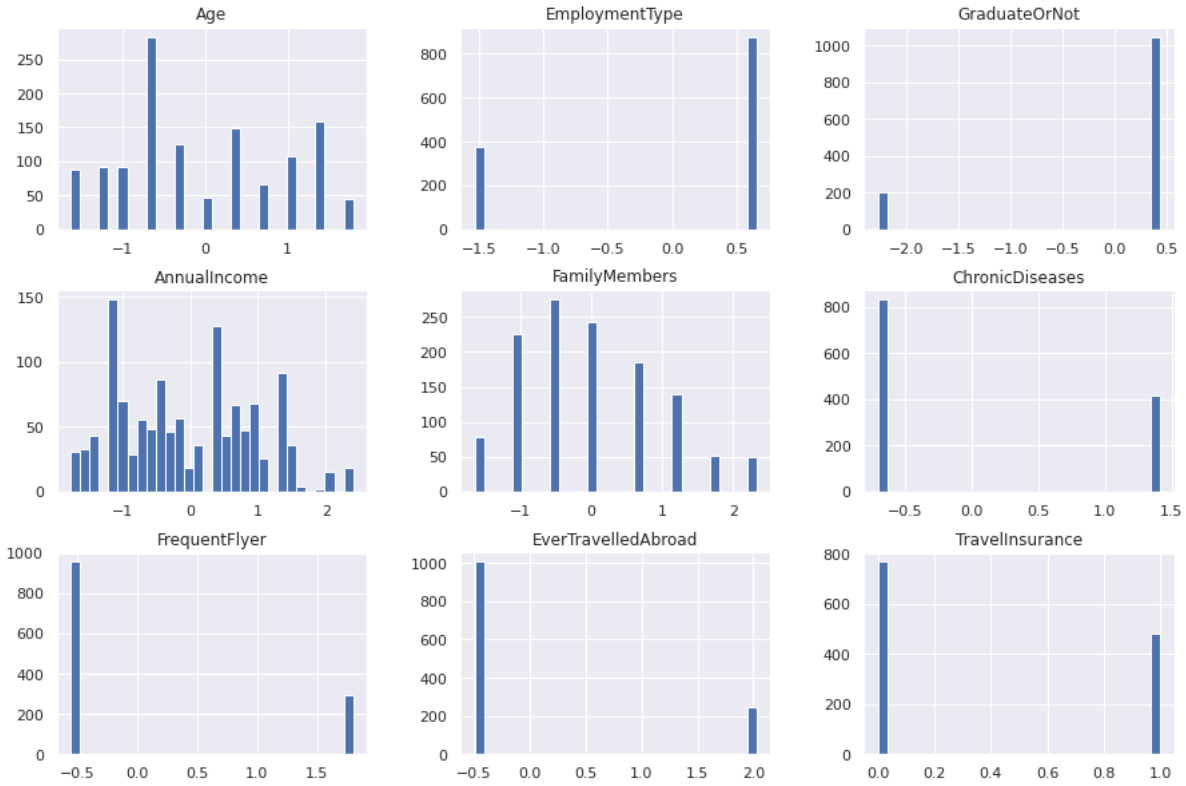

Fig.3 Histogram of Predictor variables

Chi-Square Test for Correlation Analysis of Categorical Variables

The Chi-Square Test was applied to examine the association between the binary category variables in this dataset.

Null hypothesis H0: The categorical variables do not correlate with one another.

Alternate Hypothesis H1: The categorical variables do correlate with one another.

Chi-Square Test Inferences:

If the p-value is higher than 0.05, we are unable to reject the null hypothesis and conclude that the categorical variable is not correlated with the target variable. We reject the null hypothesis and determine that the categorical variable is correlated with the target variable if the p-value obtained is less than 0.05.

The Chi-Square Test's finding:

As the p-value for each variable is higher than 0.05, we are unable to rule out the NULL hypothesis and concluded that there is no correlation between any of the categorical variables and the target variable.

Numeric Variables

After performing the data pre-processing steps, there were 1249 instances, and 3 numeric variables present in the dataset. We have to perform Scaling before proceeding. As observed from the figure, through the univariate exploratory data analysis, it was observed that ‘FamilyMembers’ variables are skewed towards the left, indicating that as expected, most of the real-world data is skewed in

nature and not normally distributed. The variable "AnnualIncome' is multi-modal in nature, while the feature ‘Age’ is bi-modal in nature.

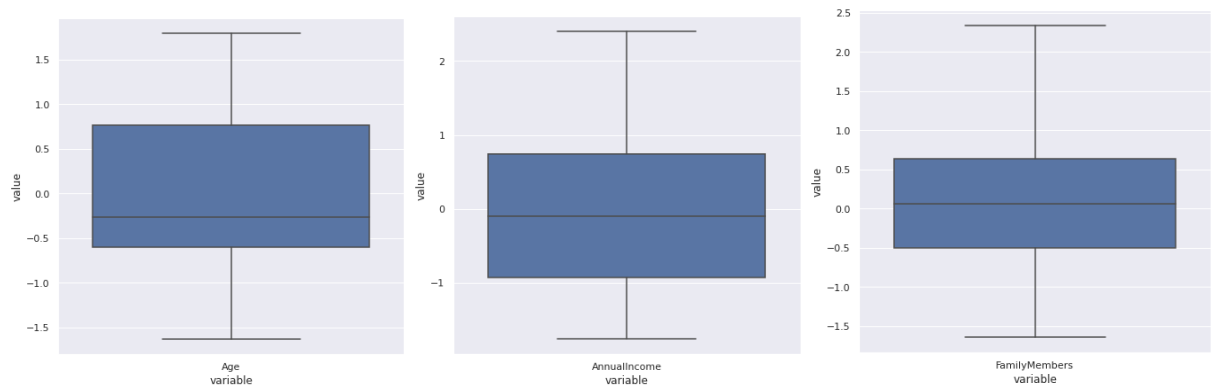

Fig.4 Box plot for Numerical predictor variables

Point Biserial Test for Correlation Analysis of Numeric and Categorical Target Variable
The Point Biserial Test is used for measuring the relationship between the binary target variable and the continuous variables of this dataset.
Null hypothesis H0: continuous variable and target variable do not correlate them.
Alternate Hypothesis H1: continuous variable and target variable correlate them.

Point Biserial Test Inferences:
If the p-value is higher than 0.05, we are unable to reject the null hypothesis and conclude that the categorical variable is not correlated with the target variable. We reject the null hypothesis and determine that the categorical variable is not correlated with the target variable as the p-value obtained is greater than 0.05.

The Point Biserial Test’s finding:
As the AnnualIncome has a p-value of more than 0.05, we are unable to reject the null hypothesis and conclude that there is no correlation between these numerical variables and the category target variable. As the p-value is less than 0.05 for all the other numerical variables, we reject the null hypothesis and conclude that these numerical variables and the categorical target variable are correlated.

Dimension Reduction and Variable Selection:

As none of the predictor variables are correlated with each other. There is no dimension reduction. All the variables are selected.

We have also implemented PCA dimension reduction, but it did not yield the expected results.

Data Mining Models and Performance Evaluation:

In our predictions, we used a total of six models. The models were logistic regression, decision tree, SVM, KNN, Naïve Bayes, Ada boost classifier, neural networks, and random forest regression. We split the data into 25% test data and 75% train data.

1. Logistic Regression :

Logistic regression is a popular statistical and machine learning algorithm used for binary classification tasks. It is a type of regression analysis that is used to predict the probability of a binary target variable based on one or more predictor variables.

Advantages:
• Logistic regression is a simple algorithm that is easy to implement and understand, making it a popular choice for many classification problems.
• Logistic regression performs well when the data is linearly separable, making it suitable for many real-world datasets.

Disadvantages:
• Logistic regression is limited to binary classification tasks and cannot be used for multi-class classification without modification.
• Logistic regression assumes a linear relationship between the predictor variables and the log-odds of the target variable, which may not always be true in real-world datasets.

Performance Evaluation:

A grid search was performed to obtain the best metrics with the highest Accuracy of 69%. The sensitivity value of 71% indicates that it was able to correctly classify the True Positives, or prediction of getting travel insurance. The specificity value of 90% indicates that it was able to correctly classify the True Negatives, or the legitimate instances A low FI-score of 48% indicates high False Positives and high False Negatives, hence incorrectly identifying predictions. The ROC curve with an AUC value of 0.64, indicating nearly perfect discrimination between getting travel insurance and not.

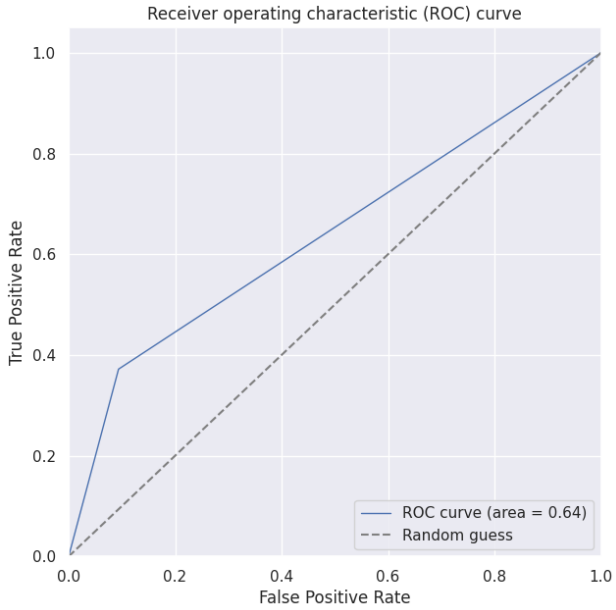

Fig.5 ROC Curve for Logistic Regression Classifier

Accuracy	Validation Error	Sensitivity	Specificity	F1 score	Classification Matrix
0.69	0.30	0.71	0.90	0.48	[[174 18] [76 45]]

2. KNN:

KNN (K-Nearest Neighbors) is a simple but powerful machine-learning algorithm used for classification and regression tasks. It works by finding the K closest data points in the training set to a given test data point and use their labels to predict the label of the test data point.

Advantages:
• KNN is a non-parametric algorithm, which means it does not make any assumptions about the distribution of the data, making it more flexible than some other algorithms.
• KNN can be used with any distance metric, allowing it to handle a wide range of data types.

Disadvantages:
• KNN is sensitive to the presence of noisy or irrelevant features in the data, which can cause it to make incorrect predictions.
• KNN does not perform well with high-dimensional data, as the distance between data points become less meaningful in high-dimensional space (known as the "curse of dimensionality").

Performance Evaluation:
A grid search was performed to obtain the best metrics neighbors=21, with the highest Accuracy of 76%. The sensitivity value of 84% indicates that it was able to correctly classify the True Positives, or prediction of getting travel insurance. The specificity value of 94% indicates that it was able to correctly classify the True Negatives, or the legitimate instances A low FI-score of 61% indicates high False Positives and high False Negatives, hence incorrectly identifying predictions. The ROC curve with an AUC value of 0.71, indicating nearly perfect discrimination between getting travel insurance and not.

Accuracy	Validation Error	Sensitivity	Specificity	F1 score	Classification Matrix
0.76	0.23	0.84	0.94	0.61	[[181 11] [63 58]]

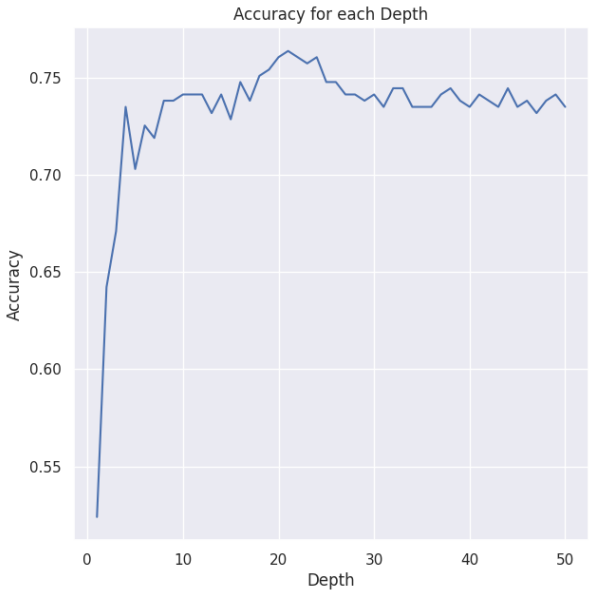

Fig.6 - Accuracy for each neighbor metric

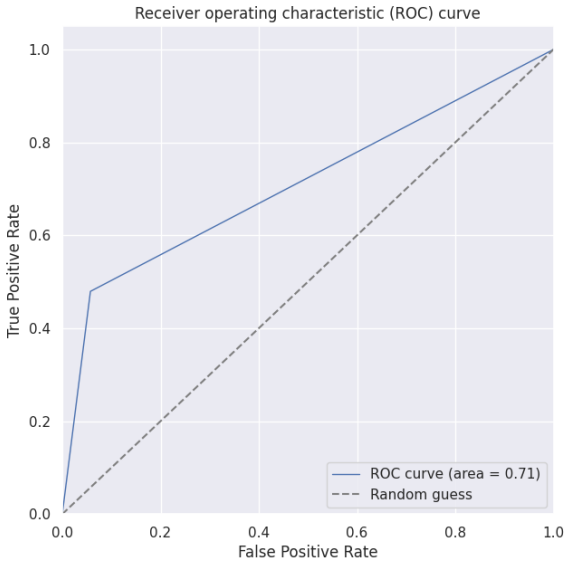

Fig.7 ROC Curve for KNN Classifier

3. Gaussian Naïve Bayes:

Naive Bayes is a probabilistic machine learning algorithm that is based on the Bayes theorem. It is used for classification and prediction tasks. Naive Bayes assumes that the features in the data are independent of each other.

Advantages

• Naive Bayes is fast and scalable, making it useful for large datasets.
• Naive Bayes can handle irrelevant features well because it ignores them, unlike other algorithms that may be negatively impacted by irrelevant features.

Disadvantages

• The independence assumption may not hold in many real-world scenarios, leading to suboptimal results.
• Naive Bayes can be sensitive to the input data distribution, particularly when it has high variance or is skewed.

Performance Evaluation:

A grid search was performed to obtain the best metrics, with the highest Accuracy of 67%. The sensitivity value of 62% indicates that it was able to correctly classify the True Positives, or prediction of getting travel insurance. The specificity value of 84% indicates that it was able to correctly classify the True Negatives, or the legitimate instances A low FI-score of 48% indicates high False Positives and high False Negatives, hence incorrectly identifying predictions. The ROC curve with an AUC value of 0.62, indicates nearly perfect discrimination between getting travel insurance or not.

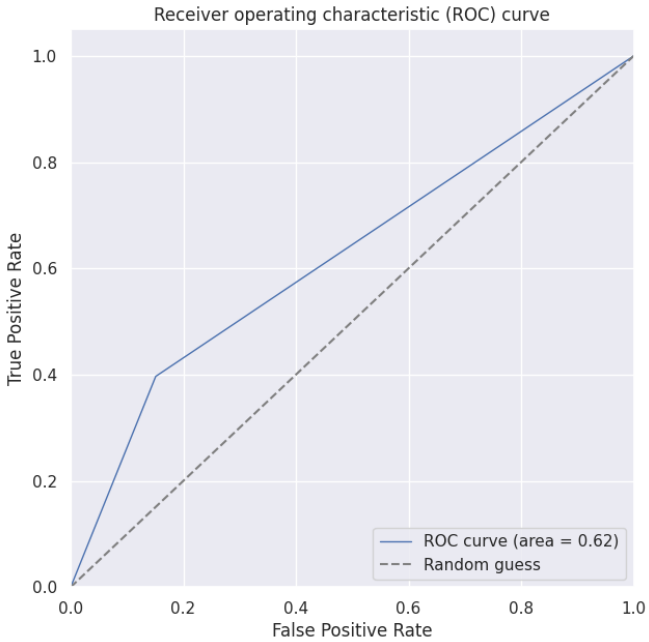

Fig.8 ROC Curve for Gaussian NB Classifier

Accuracy	Validation Error	Sensitivity	Specificity	F1 score	Classification Matrix
0.67	0.32	0.62	0.84	0.48	[[163 29] [73 48]]

4. SVM:

Support Vector Machines (SVM) is a popular classification algorithm that tries to find the best hyperplane that separates the data into different classes.

Advantages:

• SVM can handle both linear and non-linear classification problems using different types of kernels.

• SVM can also handle datasets with noisy data by using a soft-margin classifier.

Disadvantages:

• SVM can be sensitive to the choice of kernel and the value of the regularization parameter, and it may not perform well if the data is not well-separated or contains outliers.

• SVM can be affected by imbalanced datasets, where one class has much fewer samples then the other.

Performance Evaluation:

A grid search was performed to obtain the best metrics of kernel = poly, with the highest Accuracy of 75%. The sensitivity value of 83% indicates that it was able to correctly classify the True Positives or predictions of getting travel insurance. The specificity value of 94% indicates that it

was able to correctly classify the True Negatives, or the legitimate instances A low FI-score of 58% indicates high False Positives and high False Negatives, hence incorrectly identifying predictions. The ROC curve with an AUC value of 0.69, indicates nearly perfect discrimination between getting travel insurance and not.

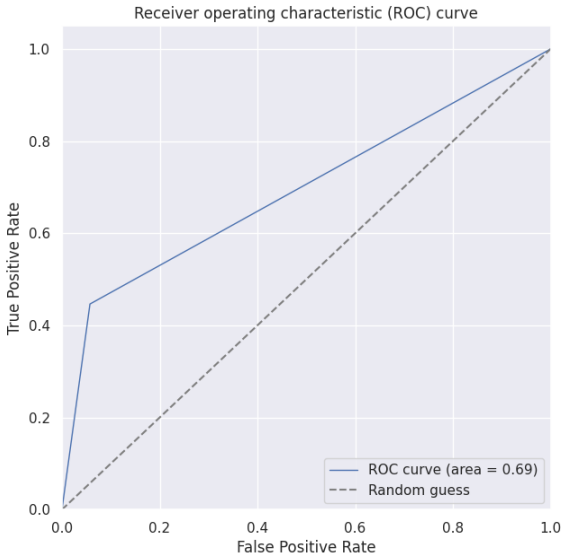

Fig.9 ROC Curve for SVM Classifier

Accuracy	Validation Error	Sensitivity	Specificity	F1 score	Classification Matrix
0.75	0.24	0.83	0.94	0.58	[[176 16] [63 58]]

5. Decision Tree :

A Decision Tree is a popular machine-learning algorithm that is widely used for classification and regression tasks. It is a supervised learning algorithm that learns a decision tree from the training data and then uses the tree to make predictions on new data. The decision tree is a tree-like model where each internal node represents a test on an attribute, each branch represents an the outcome of the test, and each leaf node represents a class label or a numerical value.

Advantages:
• The decision tree algorithm can handle both categorical and numerical data, which makes it is versatile and useful for a wide range of applications.
• The decision tree algorithm can handle missing data by splitting the data along the available attributes, which makes it useful for datasets with missing values.

Disadvantages:
• Decision tree is prone to overfitting, especially when the tree is deep and complex.
• Decision tree is sensitive to small variations in the data, which means that small changes in the data can lead to a completely different tree.

Performance Evaluation:
A grid search was performed to obtain the best metrics of max depth=3, random state=50, with the highest Accuracy of 77%. The sensitivity value of 89% indicates that it was able to

correctly classify the True Positives, or predictions of getting travel insurance. The specificity value of 96% indicates that it was able to correctly classify the True Negatives or the legitimate instances A low FI-score of 62% indicates high False Positives and high False Negatives, hence incorrectly identifying predictions. The ROC curve with an AUC value of 0.72, indicating nearly perfect discrimination between getting travel insurance or not.

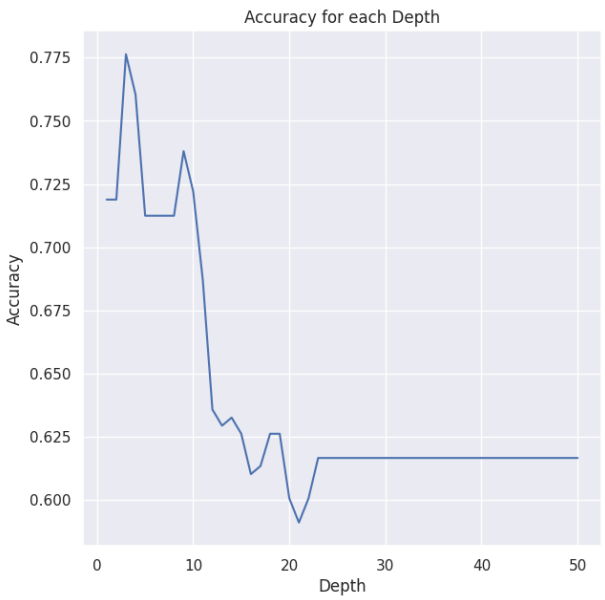

Fig.10 - Accuracy for each depth metric

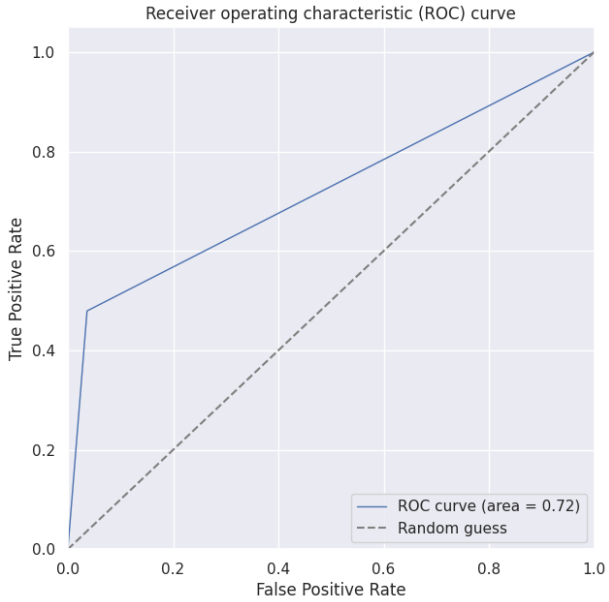

Fig.11 ROC Curve for Decision Tree Classifier

Accuracy	Validation Error	Sensitivity	Specificity	F1 score	Classification Matrix
0.77	0.22	0.89	0.96	0.62	[[185 7] [63 58]]

6. Random Forest:

Random Forest is a supervised machine-learning algorithm that belongs to the ensemble learning family. It is a collection of decision trees, where each tree is constructed using a random subset of features and a random subset of the training data. The final prediction is then made by aggregating the predictions of all the trees.

Advantages:
• Random Forest is less sensitive to outliers and noise compared to other algorithms like Decision Trees, making them more robust.
• Random Forest provides a measure of feature importance, which can help in understanding which features are most relevant for making predictions.

Disadvantages:
• Although Random Forest is less prone to overfitting than a single decision tree, it can still be overfit to noisy or irrelevant features, leading to poor performance on new data.
• Random Forest has several hyperparameters that need to be tuned, such as the number of trees, the depth of the trees, and the size of the random feature subsets. Tuning these

hyperparameters can be time-consuming and require careful experimentation to find the best values.

Performance Evaluation:
A grid search was performed to obtain the best metrics of max depth=17, and estimators=6, with the highest Accuracy of 77%. The sensitivity value of 89% indicates that it was able to correctly classify the True Positives or predictions of getting travel insurance. The specificity value of 96% indicates that it was able to correctly classify the True Negatives, or the legitimate instances A low FI-score of 62 % indicates high False Positives and high False Negatives, hence incorrectly identifying predictions. The ROC curve with an AUC value of 0.72, indicates nearly perfect discrimination between getting travel insurance and not.

Accuracy	Validation Error	Sensitivity	Specificity	F1 score	Classification Matrix
0.77	0.22	0.89	0.96	0.62	[[185 7] [63 58]]

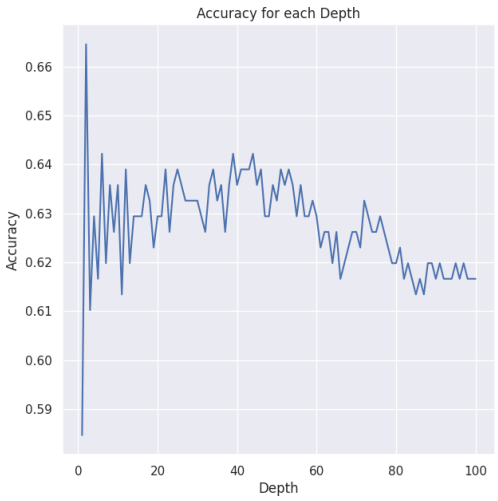

Fig.12 - Accuracy for each estimator's metric

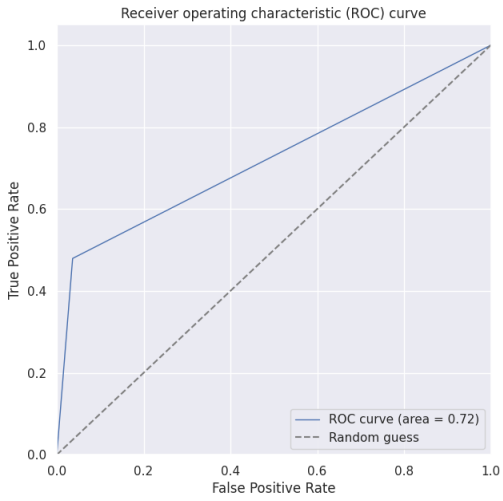

Fig.13 ROC Curve for Random Forest Classifier

7. Ada Boost Classifier:

A grid search was performed to obtain the best metrics of learning rate=1, and estimators=17, with the highest Accuracy of 76%. The sensitivity value of 85% indicates that it was able to correctly classify the True Positives or predictions of getting travel insurance. The specificity value of 95% indicates that it was able to correctly classify the True Negatives, or the legitimate instances A low FI-score of 59% indicates high False Positives and high False Negatives, hence incorrectly identifying predictions. The ROC curve with an AUC value of 0.70, indicates nearly perfect discrimination between getting travel insurance and not.

Accuracy	Validation Error	Sensitivity	Specificity	F1 score	Classification Matrix
0.76	0.23	0.85	0.95	0.59	[[184 8] [80 41]]

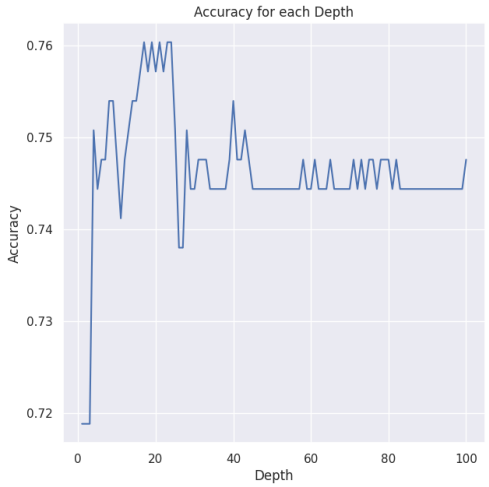

Fig.14- Accuracy for each estimator metric

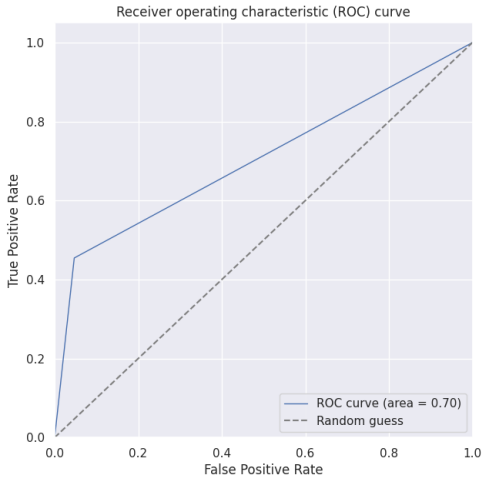

Fig.15 ROC Curve for AdaBoost Classifier

8. Neural Networks:

A grid search was performed to obtain the best metrics of solver='lbfgs', alpha=1e-5,hidden_layer_sizes=(5,2), random_state=1 , with the highest Accuracy of 72%. The sensitivity value of 75% indicates that it was able to correctly classify the True Positives or predictions of getting travel insurance. The specificity value of 90% indicates that it was able to correctly classify the True Negatives, or the legitimate instances A low FI-score of 55% indicates high False Positives and high False Negatives, hence incorrectly identifying predictions. The ROC curve with an AUC value of 0.68, indicating nearly perfect discrimination between getting travel insurance or not.

Accuracy	Validation Error	Sensitivity	Specificity	F1 score	Classification Matrix
0.72	0.27	0.75	0.90	0.55	[[174 18] [67 54]]

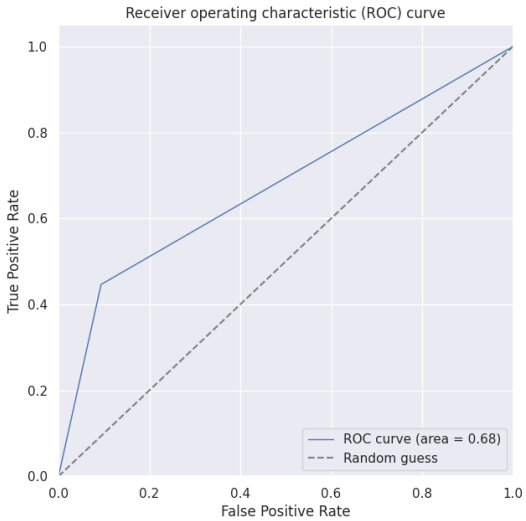

Fig.16 ROC Curve for Neural Networks Classifier

Project Summary:

We have also implemented Principal Compound Analysis on the dataset, but the results are not as good as before PCA.

Model	BEFORE PCA					AFTER PCA				
	Accuracy	Validation Error	Sensitivity	Specificity	F1 score	Accuracy	Validation Error	Sensitivity	Specificity	F1 score
Logistic Regression	0.69	0.3	0.71	0.9	0.48	0.69	0.3	0.66	0.86	0.51
Decision Tree	0.77	0.22	0.89	0.96	0.62	0.72	0.27	0.79	0.93	0.51
Random Forest Tree	0.77	0.22	0.89	0.96	0.62	0.73	0.26	0.79	0.92	0.56
Gaussian NB	0.67	0.32	0.62	0.84	0.48	0.72	0.27	0.75	0.9	0.55
SVM	0.75	0.24	0.83	0.94	0.58	0.74	0.25	0.8	0.92	0.58
Neural Networks	0.72	0.27	0.75	0.9	0.55	0.75	0.24	0.82	0.93	0.59
AdaBoost Classifier	0.76	0.23	0.85	0.95	0.59	0.31	0.68	0.32	0.05	0.45
KNN	0.76	0.23	0.84	0.94	0.61	0.75	0.24	0.83	0.94	0.58

Project Results:

Model	Accuracy	Validation Error	Sensitivity	Specificity	F1 score
Logistic Regression	0.69	0.30	0.71	0.90	0.48
Decision Tree	0.77	0.22	0.89	0.96	0.62
Random Forest Tree	0.77	0.22	0.89	0.96	0.62
Gaussian NB	0.67	0.32	0.62	0.84	0.48
SVM	0.75	0.24	0.83	0.94	0.58
Neural Networks	0.72	0.27	0.75	0.90	0.55
AdaBoost Classifier	0.76	0.23	0.85	0.95	0.59
KNN	0.76	0.23	0.84	0.94	0.61

The Decision Tree Classifier model and Random Forest Classifier has the highest overall accuracy of 77% and the lowest validation error of 22% in the classification of travel insurance instances. It also has an F-1 score of 62%. It has the highest sensitivity value of 89%, which implies that the Decision Tree Classifier and Random Forest Classifier are the best model among the other models in classifying the True Positives, which in this case are the travel insurance instances.

Impacts:

The impact of predicting whether a person will buy travel insurance or not can be significant for various stakeholders involved in the travel and insurance industries.

- Improved marketing strategies: By predicting the likelihood of a person buying travel insurance, travel companies and insurers can tailor their marketing strategies to better reach and convert potential customers. This could result in more effective and efficient marketing campaigns, leading to increased sales.

- Increased revenue: By accurately predicting who is likely to buy travel insurance, insurers can focus their efforts on high-probability customers and offer them more personalized coverage options. This could lead to increased sales and revenue for insurance companies.

- More relevant insurance products: By predicting the preferences and buying behavior of potential customers, insurance companies can offer more relevant insurance products that better meet the needs and expectations of customers. This could lead to more satisfied customers and higher customer retention rates.

- Enhanced customer experience: With more personalized coverage options and more relevant insurance products, customers are likely to have a better experience when buying travel insurance. This could lead to increased customer loyalty and repeat business.

- Reduced risk and costs: By accurately predicting who is likely to buy travel insurance, insurance companies can better manage their risk and reduce the cost of underwriting policies. This could lead to improved profitability and financial stability for insurance companies.

References:

1. Han, H., & Huang, Y. (2017). Predicting consumer behavior on travel insurance purchasing: Evidence from online and offline channels. Journal of Travel Research, 56(7), 877-891.

2. Kim, M. J., & Chung, N. (2011). Understanding the factors influencing purchase of travel insurance. Journal of Travel Research, 50(2), 205-216.

3. Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The impact of online user reviews on hotel room sales. International Journal of Hospitality Management, 30(3), 525-531.

4. Wang, D., & Li, X. (2019). Predicting travel insurance purchase: An exploratory study of Chinese outbound tourists. Journal of Travel Research, 58(5), 828-844.