Predicting Customer Interest in Enhanced Travel Insurance with COVID Cover

Venkatesh Bollineni – 800792618

SIUE

1. Abstract:

This project addresses the challenge faced by a travel company in predicting customer interest in a new insurance package, which includes COVID cover. Leveraging a dataset comprising information on nearly 2000 previous customers, the objective is to develop a predictive model for identifying potential buyers. The dataset, extracted from the 2019 sales performance, undergoes meticulous data cleaning and exploratory analysis. Recognizing the nature of the problem as a classification task, the study chooses the best model, gradient boosting from various models, as the model of choice. Through cross-validation and regularization techniques, the model achieves an optimal test error rate of 16.9%, accuracy of 83.1%, and recall of 90.8%. The final model is tested on a separate dataset of 100 unseen observations, revealing that 81 employees will not prefer TravelInsurance, whereas only 19 employees chose to have TravelInsurance, demonstrating its robustness in predicting customer interest. The project showcases a comprehensive approach to leveraging historical data for predictive modelling in the context of insurance package adoption.

2. Introduction:

In the domain of customer-centric services, a travel company is on a mission to enrich its offerings with a novel travel insurance package, together with COVID coverage. To refine its marketing strategy and meet the distinct needs of its customers, the company is leveraging data extracted from interactions with nearly 2000 clients. The data, gathered from 2019 sales records, provides a close look at how people responded to the new insurance package.

The primary objective of this analysis is to recognize the patterns within the dataset that explains the factors influencing customer interest in the travel insurance offering. In the exploration is a binary target variable, "TravelInsurance," denoting whether a customer opted to purchase the insurance package or not. Complementing this, a set of predictor variables explore into key aspects of the customer's profile, details such as age, employment type, and educational background and annual income. Crucial familial considerations, health status, travel history, and engagement with air travel services further expands the dataset.

The focal point of this analysis is the application of classification methods. Leveraging this approach, the goal is to build a predictive model capable of perceptive customers likelihood to buy travel insurance using past information. The predictor variables serve as vital inputs into the model, contributing to the formulation of a comprehensive understanding of customer behaviour. The exploration within this dataset focus advertising on specific ways, thereby optimizing the promotion and uptake of the new travel insurance package in a customer-centric method.

3. Model Specification:

The process of model specification is a critical step in building an effective predictive model that matches with the objectives of the analysis. In this study, which focuses on predicting customer interest in a travel insurance package, the choice of model moves around a classification framework. The target variable, "TravelInsurance," assists as the binary outcome, signifying whether a customer opted to purchase the insurance package or not.

The candidate models considered for this analysis are primarily from the domain of classification algorithms, given the nature of the predictive task. The suitability of the following models was assessed:

3.1 Random Forest:

Rationale: Random Forest is an ensemble method known for its robustness and ability to handle complex relationships in data. It is particularly suitable when dealing with a mix of categorical and numerical predictor variables. The core characteristic of bagging with the diversity from multiple decision trees, makes it as a strong performer for capturing complex customer behaviour patterns. Additionally, Random Forest incorporates the selection of a random subset of features at each node during the tree-building process. This feature selection mechanism not only enhances the model's predictive performance but also helps mitigate the risk of overfitting by reducing the correlation between individual trees in the ensemble. This flexibility in choosing subsets of features contributes to the algorithm's agility, making it well-suited for a wide range of datasets and improving its generalization to new and unseen data.

3.2 Gradient Boosting:

Rationale: Gradient Boosting is an ensemble technique that builds decision trees sequentially, with each tree correcting the errors of its predecessor. This method is adept at capturing non-linear relationships and interactions within the data. Considering the various factors and detailed connections among predictor variables. Gradient Boosting involves fitting sub-models, typically decision trees, to residuals. Its iterative nature comprises three key elements: a loss function for optimization, a weak learner for predictions, and an additive model to minimize the loss function.

During boosting, each tree is fitted using the current residual, not the outcome Y. This allows for relatively small trees with limited splits. The introduction of a shrinkage parameter, often a small positive number like 0.01 or 0.001, slows down the updating process, contributing to a controlled and robust learning process. This, combined with the iterative nature of boosting, mitigates the risk of overfitting, enhancing the model's generalization performance on new data.

3.3 Support Vector Machines (SVM):

Rationale: SVM is a powerful classification algorithm that performs well in scenarios where decision boundaries are not necessarily linear. By mapping features into a higher-dimensional space, SVM seeks to find an optimal hyperplane that maximally separates different classes. Given the complexity in the relationship between predictor variables and customer interest, SVM provides a detailed framework for capturing non-linear patterns.

Linear Kernel, suitable for linearly separable data, the linear kernel maps features into a higher-dimensional space without introducing non-linearity in the decision boundaries. *Radial Basis Function (RBF) Kernel,* The RBF kernel introduces non-linearity by considering the similarity between data points in the original feature space. This enables SVM to create complex decision boundaries, for capturing non-linear patterns. *Polynomial Kernel,* the polynomial kernel introduces non-linearity by applying polynomial transformations to the original features. It allows SVM to model more detailed relationships by considering polynomial decision boundaries of varying degrees.

3.4 Logistic Regression:

Rationale: Logistic Regression is a reliable and easy-to-understand model used for tasks like deciding between two options. It helps us see how different things affect the chances of a positive result, like customers being interested in travel insurance. This model is great when we want to know the impact of each factor on predictions, making it simple for people to understand. Logistic Regression is also good at finding the right balance between making the model too simple or too complicated, which is important for making accurate predictions without being too complex.

3.5 Linear Discriminant Analysis (LDA):

Rationale: Linear Discriminant Analysis (LDA) is a method used to classify things by finding the best combination of factors that make them different. It looks for a way to draw lines between different groups in the data, making sure that the differences between groups are as big as possible.LDA is good when we can use straight lines to separate different groups in the data. It works well when the reasons for differences between groups can be explained by simple, straight relationships between the factors we are looking at. LDA focuses on making the groups as different as possible while keeping things similar within each group.

It is impossible to choose a model by looking at their characteristics unless we fit the model. So, the subsequent sections will delve into the fitting, tuning, regularization, diagnostics, evaluation, and comparison of these models to ascertain their effectiveness in predicting customer interest in the travel insurance package.

4. Fitting and Diagnostics:

As mentioned above it was chosen five classification models and are fitted for the trained data set (70% of observations), tested with the test data (30% of observations) set and chosen best model depending on the various important statistical metrics. Mainly focused on metrics like test error rate on test set, Recall and Accuracy in improving and choosing the best statistical model. Though these chosen metrics are completely depends on the requirements of the prediction and varies based on the nature of the problem, here we have chosen randomly based on previous works. Though we have limited our preferred metrics, we have evaluated other crucial metrics such as precision and F1 score, plotted the ROC curves as well wherever significant.

4.1 Random Forest Model:

The training process commenced with the training dataset, where the response variable, TravelInsurance, was predicted based on a set of predictor variables. The parameter ntree was set to 1000 to determine the number of trees in the forest, influencing the model's complexity and potential for overfitting. To identify significant variables in model fitting, a graph *(figure 1)* was plotted for variable importance.

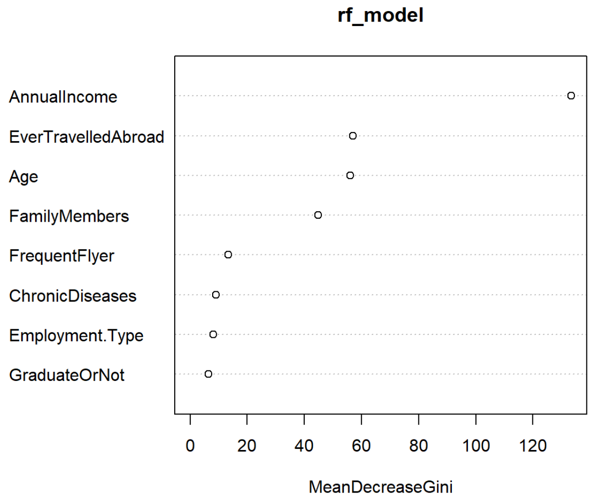

Figure 1: Graph plotted between MeanDecreaseGini and predictor variables, the graph indicates AnnualIncome is the most important variable following by EverTravelledAbroad and Age.

4.1.1 Applying Validation technique and tuning:

The model fitting process was further enriched through k-fold cross-validation with 10 folds using the trainControl method. This step robustly assessed the model's performance on different subsets of the training data. Cross-validated results for various mtry values indicated that the model with mtry = 3 was chosen. Utilizing the trained data with the cross-validation model, the top 3 predictors were employed to further train the model, with the aim of improving metrics such as test error rate, Recall and Accuracy.

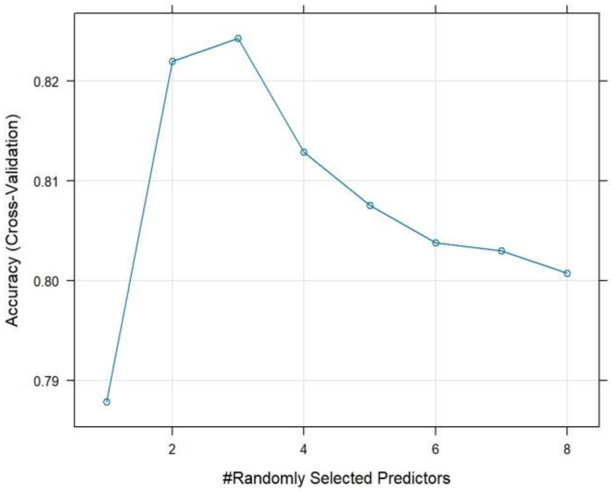

Figure 2: The plot shows the selection of number of predictors during cross validation and when using 3 predictors the accuracy was yielded.

Despite selecting mtry = 3 (only top 3 predictors in fitting the model) as the optimal parameter, the test error rate 17.28% and accuracy 82.71% exhibited no improvement. A confusion matrix was plotted, notably, there was a decrease in Recall from 91.26% to 87.68%, indicating a reduction in the model's ability to correctly identify all instances of interest in TravelInsurance. However, precision improved from 56.93% to 59.90%, signifying enhanced accuracy in positive predictions, although at the potential cost of false positives. The F1 score, a balanced metric considering precision and recall, experienced a slight increase from 0.7012 to 0.7117. While the model demonstrated overall good performance, a precision value of 0.569 suggested potential false positives, emphasizing the ongoing need for a delicate balance between precision and recall.

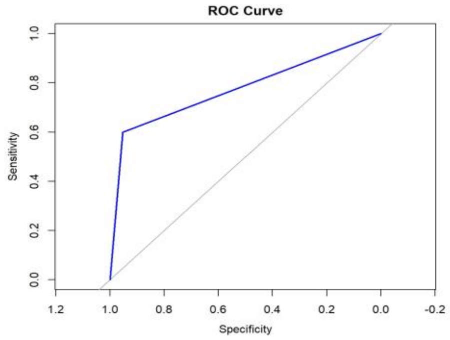

Figure 3: AUC (Area under the curve), with a value of 0.7762173, quantifies the model's overall ability to discriminate between positive and negative classes.

4.2 Gradient boosting model:

Gbm (Gradient Boosting Machine) algorithm was initially trained with 100 trees. The model aimed to predict the likelihood of individuals purchasing travel insurance. The achieved accuracy on the test set is 82.89%, indicating the overall correctness of predictions with test error rate of 17.11%. However, our primary focus lies in improving the model's ability to capture positive cases, as reflected in the Recall metric, which currently stands at 90.08%. High Recall is crucial in this context, as it minimizes the risk of failing to identify potential customers interested in travel insurance, reducing false negatives. The test error rate and Precision, Recall, and F1 Score metrics further provide a comprehensive evaluation of the model's performance. The confusion matrix reveals 84 false positives and 13 false negatives, emphasizing the importance of refining the model to reduce both types of errors. The Area Under the Curve (AUC) value of 0.789231 and the ROC (Receiver Operating Characteristic) curve illustrate the trade-off between true positive and false positive rates.

4.2.1 Applying validation techniques and shrinkage regularization:

To enhance the model's performance, a systematic tuning process was employed. This involved cross-validated tuning using 10 folds and a grid of parameters, including the number of trees (n.trees), interaction depth, shrinkage, and minimum observations in a terminal node (n.minobsinnode). Through this comprehensive parameter search, the optimal model configuration was identified as having 150 trees, an interaction depth of 3, shrinkage of 0.05, and a minimum of 5 observations in a terminal node. The cross-validation technique ensured robust evaluation across different subsets of the training data, contributing to the selection of the most effective model configuration.

The subsequent evaluation of the tuned model on the test set revealed promising results. The recall improved to 90.8%, accuracy improved to 83.1%, and test error rate reduced to 16.9% other essential performance metrics, including precision (58.4%), F1 score (0.711), and AUC (0.776), provided a nuanced understanding of the model's strengths. Despite these positive outcomes, deficiencies were identified, including the need for a balance between precision and recall. The model's sensitivity to certain parameters and its generalization to new data were crucial considerations. This study emphasizes the iterative nature of model development, where tuning parameters and evaluating performance on distinct datasets contribute to a refined understanding of the model's predictive capabilities and potential limitations.

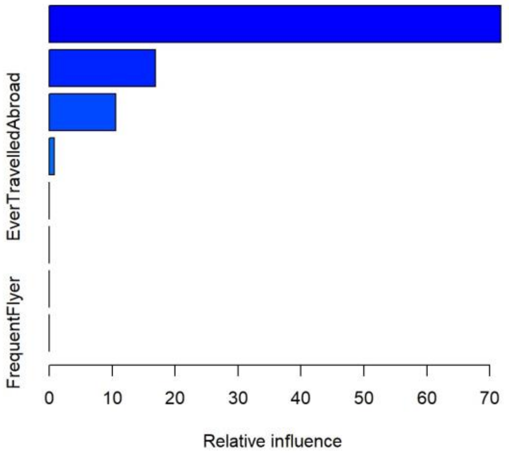

Figure 4: Gradient boosting model identified there are only 3 most important predictor variables in training the dataset.

4.3 Support Vector Machines:

4.3.1 Applied validation techniques and tuning and regularization:

Support Vector Machine (SVM) model was chosen as the next model in predicting the target variable TravelInsurance. A linear kernel SVM was trained on the training set by tuning with various cost parameters using 10-fold cross-validation and the optimal cost was found to be 0.006, resulting in test error rate of 22.22%, accuracy of 77.78%, recall of 48.02%, precision of 82.20%, and an F1 score of 0.6063.

Next, radial and polynomial kernels were explored to evaluate their impact on model performance. The radial kernel, with various values of tuning parameters, found that gamma of 0.1 and a cost of 10, achieved with a test error rate of 19.4%, an accuracy of 80.60%, recall of 57.43%, precision of 82.86%, and an F1 score of 0.6784. For the polynomial kernel, a degree of 4 and a cost of 10 were identified as optimal through cross-validated tuning, resulting in an improved results with a test error rate of 18.69%. accuracy of 81.31%, recall of 56.93%, precision of 85.82%, and an F1 score of 0.6845.

In comparing the linear, radial and polynomial kernels for the SVM model, the polynomial kernel exhibited superior performance with a lower test error rate of 18.69%, highlighting its improved accuracy in predicting travel insurance interest compared to the radial kernel and linear kernel. The polynomial kernel also demonstrated higher recall and precision showcasing a balanced ability to make accurate positive predictions while capturing a greater proportion of actual positive instances. Consequently, the F1 score for the polynomial kernel was slightly higher, emphasizing the model's enhanced balance between precision and recall, contributing to its overall improved predictive capabilities. These results highlight the effectiveness of SVM in predicting travel insurance interest, with the choice of kernel and associated parameters influencing the trade-off between precision and recall.

4.4 Logistic Regression:

In the logistic regression model, the initial model was trained using the entire set and exhibited coefficients that provided insights into the importance of different features. The model was then evaluated on a test set, yielding an accuracy of 75.66%, test error rate of 24.34%, precision of 68.82%, recall of 57.92%, and an F1 score of 62.90%. The summary of the logistic model showed a residual deviance of 1383.7 on 1311 degrees of freedom, indicating a reasonably good fit. Subsequently, a logistic regression model was constructed with cross-validated feature reduction, incorporating only "AnnualIncome," "FamilyMembers," and "Age." This reduced model achieved an accuracy of 75.49%. However, the diagnostic techniques revealed some deficiencies, as the model's performance did not surpass the initial logistic regression model, suggesting that the excluded features may contribute valuable information.

Additionally, while logistic regression is a valuable tool for binary classification, it assumes a linear relationship between predictors and the log-odds of the response. If the underlying relationship is nonlinear, the model may not capture complex patterns effectively. The feature reduction approach, while helpful in simplifying the model, might overlook potentially important predictors, impacting predictive performance. Moreover, the achieved accuracies, while reasonable, may require further consideration based on the specific requirements and consequences of false positives and false negatives in the context of travel insurance prediction.

4.4.1 Summary of logistic regression model:

```
Call:
glm(formula = TravelInsurance ~ ., family = binomial, data = train_set)

Coefficients:
 Estimate                     Std.      Error       z       value       Pr(>|z|)
(Intercept)                -4.986e+00    7.648e-01    -6.519    7.10e-11    ***
Age                            6.236e-02    2.250e-02    2.771    0.005591    **
Employment.TypePrivate Sector/Self Employed 9.366e-02 1.610e-01 0.582 0.560846
GraduateOrNotYes             -1.293e-01       1.905e-01       -0.679       0.497103
AnnualIncome               1.456e-06     2.138e-07     6.809     9.82e-12     ***
FamilyMembers               1.574e-01     4.089e-02     3.850     0.000118     ***
ChronicDiseases               1.242e-01       1.461e-01       0.850       0.395228
FrequentFlyerYes               2.091e-01       1.735e-01       1.206       0.227978
EverTravelledAbroadYes     1.904e+00    1.955e-01    9.741     <    2e-16    ***
---
Signif.  codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘  ’  1

(Dispersion    parameter    for    binomial    family    taken    to    be    1)

 Null     deviance:     1729.3     on     1319     degrees     of     freedom
Residual     deviance:     1383.7     on     1311     degrees     of     freedom
AIC:                                                                      1401.7

Number of Fisher Scoring iterations: 4
```

4.5 Linear Discriminant Analysis:

In the Linear Discriminant Analysis (LDA) analysis, two models were developed and evaluated for predicting travel insurance interest. The initial LDA model, trained on the full dataset, achieved recall of 43.07%, an accuracy of 73.54% with test error rate of 26.46%. Subsequently, precision, recall, and the F1 score were calculated, resulting in precision of 70.73%, and an F1 score of 53.54%.

4.5.1 Applied validation techniques:

To enhance the model and address overfitting, a feature-reduced LDA model was constructed using cross-validated feature selection, considering only "AnnualIncome," "FamilyMembers," and "Age." This reduced model showed an improved accuracy of 75.84%, test error rate 24.16%, recall of 53.96%, plotted a confusion matrix, which shows 321 true negatives, 44 false positives, 93 false negatives, and 109 true positives. Precision, recall, and the F1 score for the feature reduced LDA model were calculated, resulting in precision of 71.24%, and an F1 score of 0.6141.

Diagnostics techniques included the computation of precision, recall, F1 score, and the confusion matrix for both models. The F1 score provides a balanced measure that considers both precision and recall. Additionally, cross-validated feature selection was employed in the second model to mitigate overfitting.

	Predicted	
Actual	**0**	**1**
0	**321**	**44**
1	**93**	**109**

Table 1: Confusion matrix after the cross validation and feature reduction in fitting the model

4.6 Selection of final model:

The rationale behind selecting the best model lies in choosing the one with the lowest test error rate, highest accuracy, and optimal recall. A lower test error rate suggests superior performance on unseen data, while higher accuracy reflects the correct predictions relative to the total predictions. The emphasis on recall is crucial as it signifies the proportion of actual positives correctly identified. Among the five models considered, only the Gradient Boosting Method has met all these criteria.

Model	**Test error rate**	**Accuracy**	**Recall**	**Precision**	**Tuning methods**
Random Forest	17.28%	82.71 %	87.68%	59.90%	Cross validation and features reduction
Gradient Boosting	16.9%	83.1%	90.8%	58.4%	Cross validation, shrinkage, interaction depth
Support Vector Machine	18.69%	81.31%	56.93%	85.82%	Polynomial degree of 4 and cost of 10
Logistic Regression	24.34%	75.66%	57.92%	68.82%	*Tuning results are worse than basic model*
Linear Discriminant Analysis	24.16%,	75.84%	53.96%	71.24%	Cross validation and feature reduction.

Table 5: Comparison of Test error rate, Accuracy, Recall Precision metrics for all the trained models for selecting the best model.

4.7 Predicting the target variable TravelInsurance on the unseen 100 observations data using the best model (Gradient Boosting):

This project focuses on predicting customer interest in a new insurance package, including COVID coverage, utilizing the Gradient Boosting Method (GBM) with cross-validation and shrinkage regularization. The tuning process involves exploring different hyperparameter combinations, such as the number of trees, interaction depth, shrinkage, and minimum observations in a node, to optimize model performance. The best tuned GBM model achieves notable results, with a test error rate of 16.9%, accuracy of 83.1%, and a high recall of 90.8%. The model is then applied to a separate test dataset, where predictions are generated and appended to the original dataset. The final predictions categorize customers into those likely to purchase travel insurance ("Yes") and those less likely ("No"). The distribution of predictions is presented, indicating that out of 100 unseen observations, 81 are predicted as "No" and 19 as "Yes." This showcases the model's robustness in identifying potential buyers and emphasizes its practical application in predicting customer interest in the context of insurance package adoption.

```
[1] 0 0 1 1 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0
[54] 0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0
0 0 0 0 0 1 0 0 0 0
```

Binary representation (copied from the output of R programming) of the prediction of target variable "TravelInsurance" on unseen 100 observations, 0 indicates not preferred to but the Travel Insurance whereas 1 indicates interested in buying the Travel Insurance

```
 [1] "No" "No" "Yes" "Yes" "No" "No" "No" "No" "Yes" "No" "No" "Yes" "No" "No"
"No" "Yes" "No"
 [18] "No" "No" "No" "No" "No" "Yes" "No" "No" "Yes" "No" "No" "Yes" "Yes"
"No" "No" "No" "No"
 [35] "No" "No" "No" "No" "No" "No" "No" "No" "Yes" "No" "No" "No" "No" "No"
"No" "No" "Yes"
 [52] "No" "No" "No" "Yes" "No" "No" "No" "No" "No" "Yes" "Yes" "No" "No"
"Yes" "No" "No" "Yes"
 [69] "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "Yes"
"No" "No" "No"
 [86] "Yes" "No" "No" "No" "No" "No" "No" "No" "No" "No" "Yes" "No" "No" "No"
"No"
```

The above results are the preferences of 100 employees in buying the TravelInsurance predicted by the best model Gradient Boosting Model. No (81) represents not interested to buy the Travel Insurance and Yes (19) indicates interested in choosing the Travel Insurance.

5. Discussion:

5.1 Data cleaning and data exploration:

The dataset cleaning process commenced with the identification and handling of missing data and the removal of unnecessary columns, specifically serial numbers, which did not contribute to the analysis. Several insights were extracted during data exploration. Notably, a significant relationship between the target variable "TravelInsurance" and predictors such as

"Age," "AnnualIncome," "FrequentFlyer," and "EverTravelledAbroad" was observed through the evaluation of p-values using ANOVA (Analysis of Variance) tests.

An additional observation revealed that the count of individuals employed in the private sector or self-employed was 2.5 times higher than those employed in the government sector. Furthermore, individuals in the private sector or self-employed exhibited a higher average salary compared to their government-employed counterparts.

The target variable, indicating the presence or absence of travel insurance (0 for no insurance, 1 for insurance), presented a class imbalance. Specifically, the 0 class had 1206 observations, constituting 63% of the dataset, while the 1 class had 681 observations. To address this imbalance, a threshold technique was applied during the comparison of true labels.

5.2 Observations in training and testing of the model:

During the training and testing of the model, it was noticed that only features “Age”, “Annual Income”, “EverTravelledAbroad” contributed greatly to the fitting of the model, the remaining predictors were not significant and only induced noise in the model. Usually, this leads to poor metrics performance.

Picking test error rate, recall, and accuracy as key measures is a smart move. Test error rate tells us how well the model does on new data. Recall is important because it shows how good the model is at catching positive cases, like predicting customer interest. Accuracy is a common metric, giving a general idea of correctness. Using all these metrics together gives a complete look at the model's strength, how well it catches positive cases, and overall correctness. This method helps us understand the model's performance better, especially when dealing with imbalanced datasets. However, the choice of metrics depends on the specific problem, and others might prefer precision and F1 score as important measures.

Also, to handle the imbalanced data, it was chosen to adjust the threshold to improve the metrics, but this tweaking hasn’t given any significant improvements to the results.

5.3 Summary of the paper:

This project tackled the task of guessing if customers would be interested in a new insurance deal, especially for travel, using machine learning methods. We carefully cleaned and investigated data from around 2000 past customers. The main goal was to create a model that could predict who might want to buy the insurance.

Five machine learning models—Random Forest, Gradient Boosting, Support Vector Machines, Logistic Regression, and Linear Discriminant Analysis—were trained and evaluated. The training process involved tuning hyperparameters, applying cross-validation, and exploring various regularization techniques to enhance model performance. The models were assessed based on key metrics such as test error rate, accuracy, recall, precision, and the area under the ROC curve (AUC).

The Gradient Boosting Model emerged as the most promising, demonstrating an accuracy of 83.1%, a test error rate of 16.9%, and a recall of 90.8%. This model was selected as the best among the others due to its optimal performance in terms of the chosen criteria. The AUC value of 0.776 further confirmed the model's ability to discriminate between positive and negative classes.

The project showed that building a model is an ongoing process. It stressed how adjusting parameters and finding the right balance between precision and recall is crucial. It also recognized problems like assuming a linear relationship in logistic regression and the need to carefully handle how sensitive the model is.

The final section presented the predictions of the best model on unseen data, categorizing instances into those likely to purchase travel insurance and those less likely. The discussion concluded with a high-level comparison of the performance metrics across all models, reinforcing the selection of the Gradient Boosting Model as the most suitable for addressing the specific challenge posed by the travel company.

To refine our travel insurance models, we can test them on real-world data and explore alternative algorithms or parameter tweaks. But before deployment, we must prioritize understanding the "cost" of each error. Are false positives, where someone buys unnecessary insurance, worse than false negatives, leaving someone uninsured? This critical decision guides our final model choice.

Overall, the project showcased a comprehensive approach to predictive modelling, combining data preprocessing, model training, and evaluation to provide practical takeaways for the travel company in predicting customer interest in a new insurance package.