Prediction of Travel Insurance Purchase

ACTU 5841 Data Science in Finance and Insurance

Professor Yubo Wang

Po An Chen (pc2950)

Susan Wang (yw3688)

Wenyu Wu (ww2599)

Jing Zhang (jz3361)

I. Introduction

Travel insurance is a type of insurance policy that reimburses policy holders for money they lose from non-refundable deposits and payments when something goes wrong on their trip. These problems can range from lost baggage to flight delays to medical problems. The more someone spends on their trip, the more valuable travel insurance becomes. This is especially true for international trips and cruises, where travel problems could be quite expensive to solve. Since the COVID-19 pandemic has become a norm, many travel insurance plans are considering including epidemic-related covered reasons. Specifically, one tour & travels company in India is offering COVID-19 related coverage in addition to their existing travel insurance plan. The company wants to know which customers would be interested to buy it based on its database history. In this project, we will use the dataset provided by the company and aim to predict whether someone will purchase travel insurance based on various factors. We will also provide some suggestions in the end so that the company can conduct further research and improve upon our model.

II. Data Description

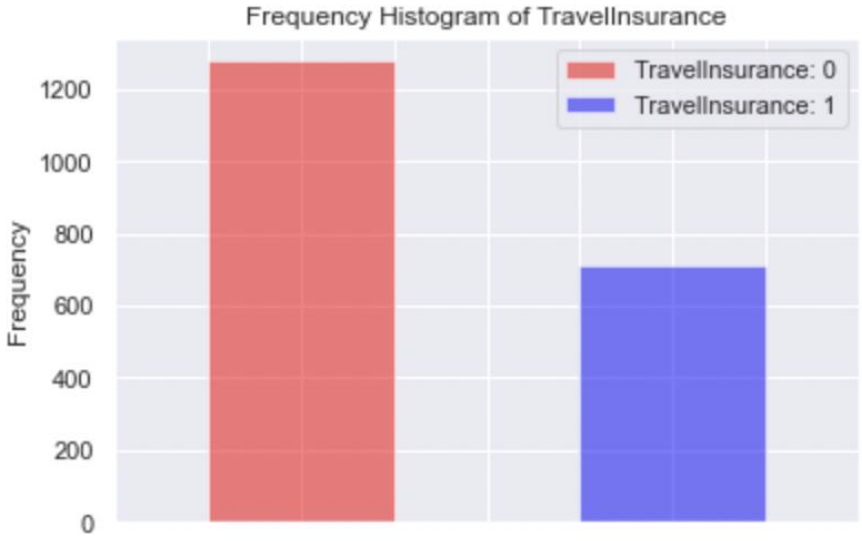
The dataset includes information about 1987 customers, whom the insurance was offered to in 2019. In addition to the target variable TravelInsurance, there are 8 predictor variables, 3 of which are numeric and 5 of which are categorical variables. The data contains no missing or seemingly erroneous values. Here is a brief description of the variables.

TravelInsurance	1 if the customer purchased travel insurance package in 2019, and 0 otherwise
Age	Age of the customer
Employment.Type	The sector in which customer is employed
GraduateOrNot	Whether the customer is college graduate or not
AnnualIncome	Annual income of the customer in Indian Rupees (rounded to nearest 50 thousand Rupees)
FamilyMembers	Number of members in customer's family
ChronicDisease	1 if the customer suffers from major disease or conditions like diabetes, high blood pressures, asthma, etc, and 0 otherwise
FrequentFlyer	Whether the customer booked at least 4 flight tickets in the past two years (2017-2019)
EverTravelledAbroad	Whether the customer has travelled to a foreign country

III. Data Visualization

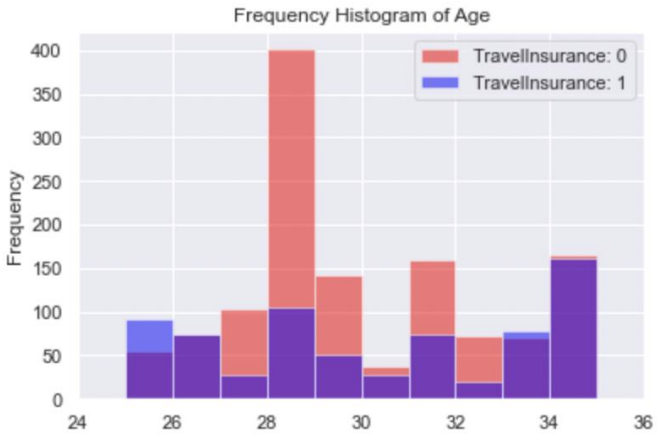
TravelInsurance

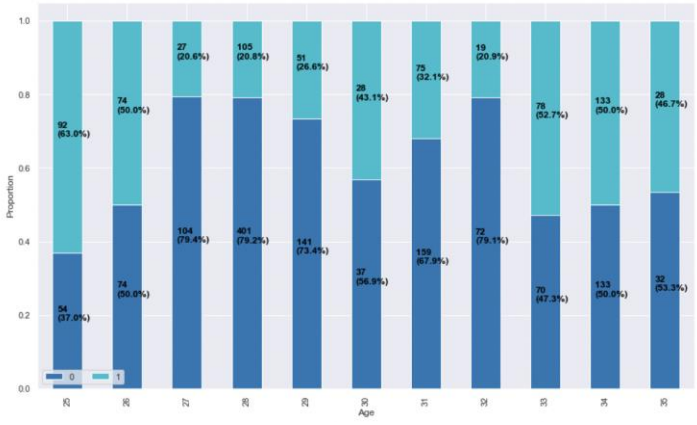
Out of 1987 observations, 710 (35.7%) purchased travel insurance and 1277 (64.3%) did not purchase travel insurance.



Age

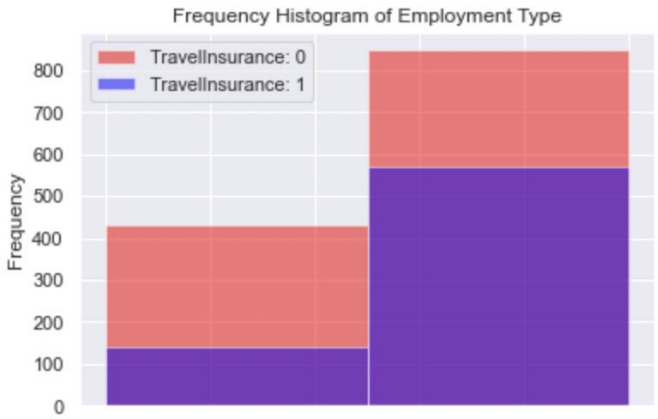
Out of 1987 observations, the minimum age is 25 and the maximum age is 35. Based on the stacked bar chart, Age seems like a useful predictor at first glance, since the proportion of people who purchased travel insurance is higher for people aged 25-26 and 33-35 compared to the rest. However, this might also be due to the fact that the proportion of people in each age group is not the same, with more people aged 27-31 compared to the rest of the age group. Another question that comes to mind is that the age range only includes younger people, and we could investigate further whether this is a representative sample of the target population.





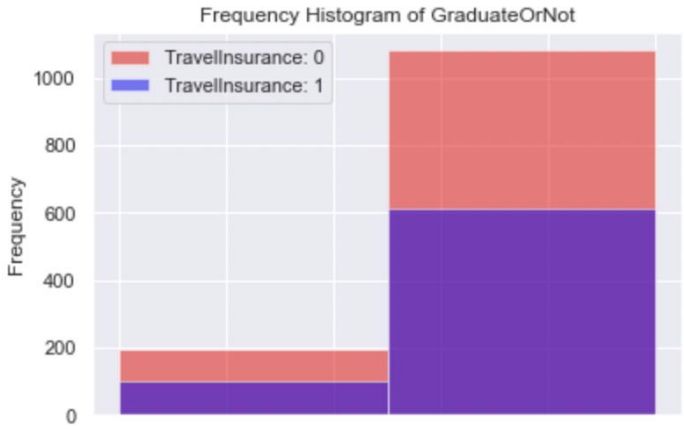
Employment Type

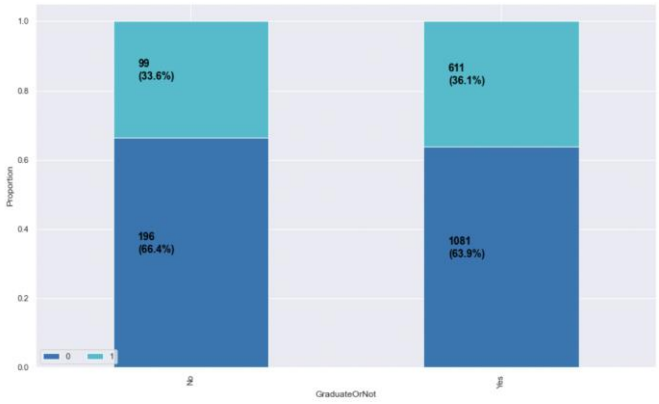
We modified the variable with 1 representing the private sector/self employed, and 0 representing employment in the government sector. Out of 1987 observations, 570 (28.7%) worked in the government sector, 140 (24.6%) of which purchased travel insurance, and 1417 (71.3%) worked in the private sector or were self employed, 570 (40.2%) of which purchased travel insurance. Though it seems that people who work in the private sector are more likely to purchase travel insurance, this pattern could also be due to the disproportion of sample in each sector. We will investigate further in the next step of our project.



GraduateOrNot

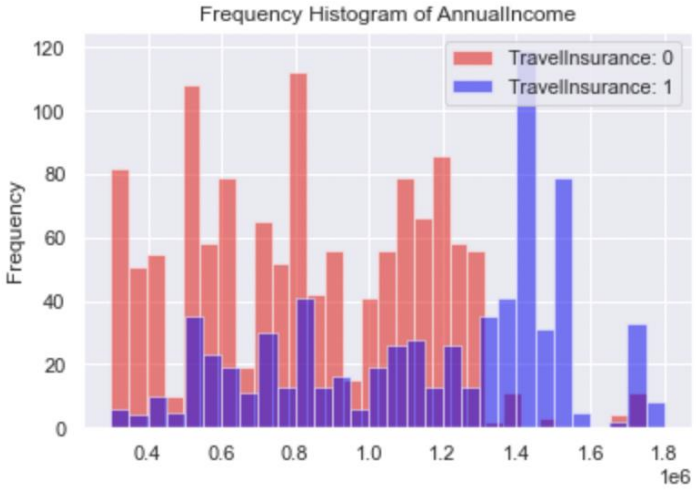
We modified the variable with 1 representing college graduates, and 0 representing not graduated from college. Out of the 1987 observations, 1692 (85.2%) graduated from college, 611 (36.1%) of which purchased travel insurance, and 295 (14.8%) did not graduate from college, 99 (33.6%) of which purchased travel insurance. At first glance, it seems that there is no significant trend between whether people graduated from college and whether they purchase travel insurance.

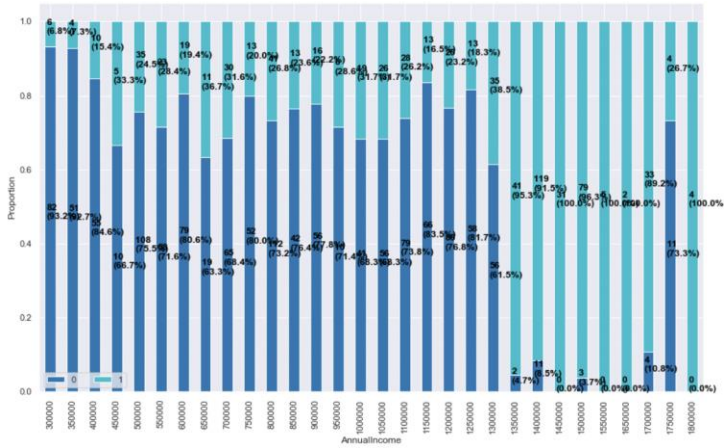




AnnualIncome (in Rupees)

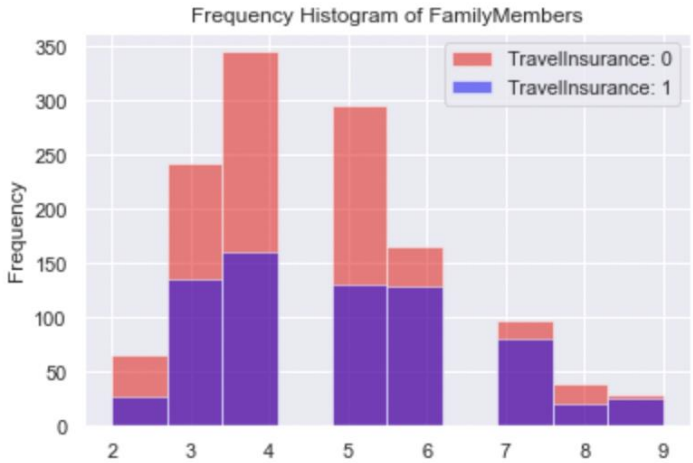
Among the 1987 observations, the minimum annual income is 300000, the maximum annual income is 1800000, and the median annual income is 900000. Based on the stacked bar chart, people with higher income are very likely to purchase travel insurance, thus at first glance AnnualIncome seems an important predictor.

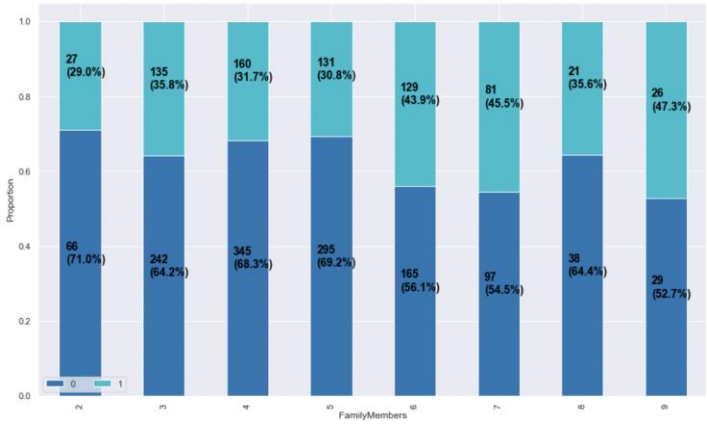




FamilyMembers

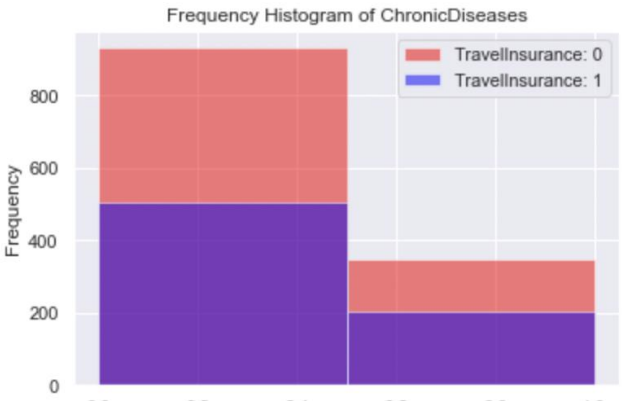
Among the 1987 observations, the minimum family members is 2, the maximum family members is 9, and the median family members is 5. Based on the stacked bar chart, people with more family members are more likely to purchase travel insurance. However, we will investigate more in the next step how significant this relationship is.

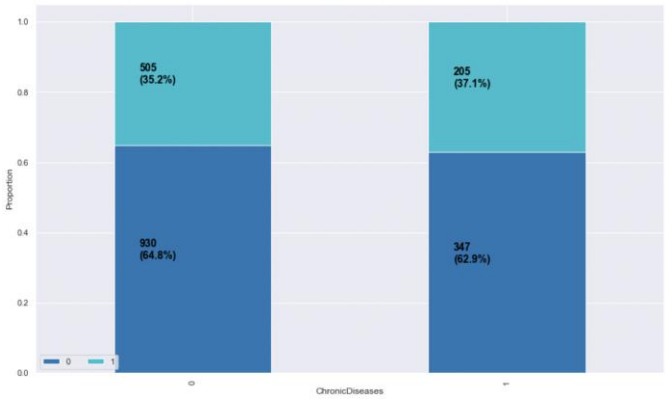




ChronicDisease

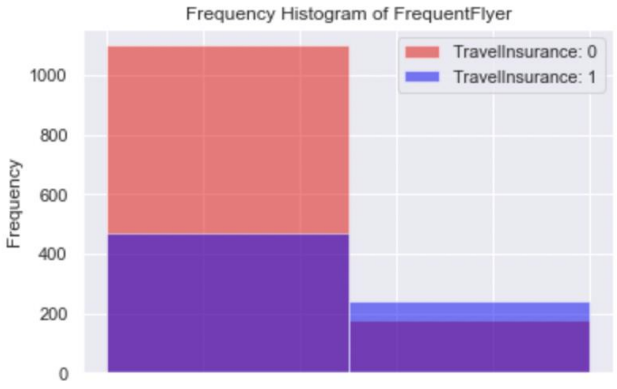
Out of 1987 observations, 552 (27.8%) had some kind of chronic diseases, 205 (37.1%) of which purchased travel insurance, and 1435 (72.2%) did not have chronic diseases, 505 (35.2%) of which purchased travel insurance. At first glance, it seems that there is no significant trend between whether people have chronic diseases and whether they purchase travel insurance.

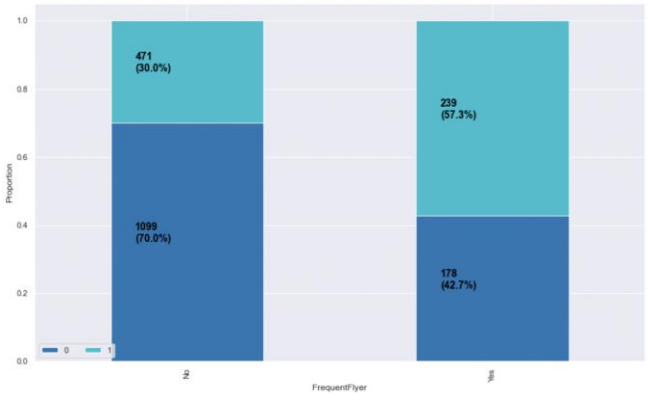




FrequentFlyer

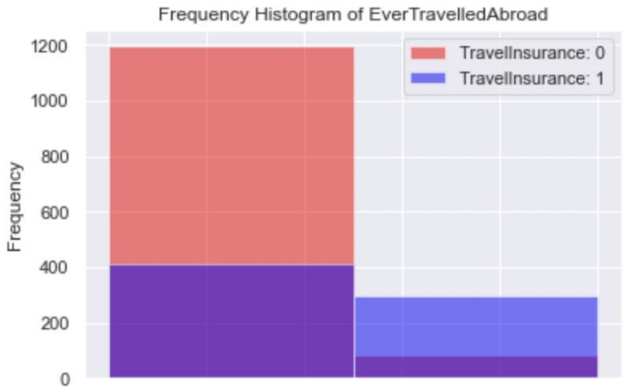
We modified the variable with 1 representing customers who booked at least 4 flights in the past two years (2017-2019), and 0 representing customers who booked less than 4 flights in the past two years. Out of 1987 observations, 417 (21%) are frequent travelers, 239 (57.3%) of which purchased travel insurance, and 1570 (79%) are not frequent travelers, 471 (30%) of which purchased travel insurance. Although the proportion of frequent flyers is small, our stacked bar chart indicates that customers are more likely to purchase travel insurance if they travel frequently, which also aligns with our intuition.

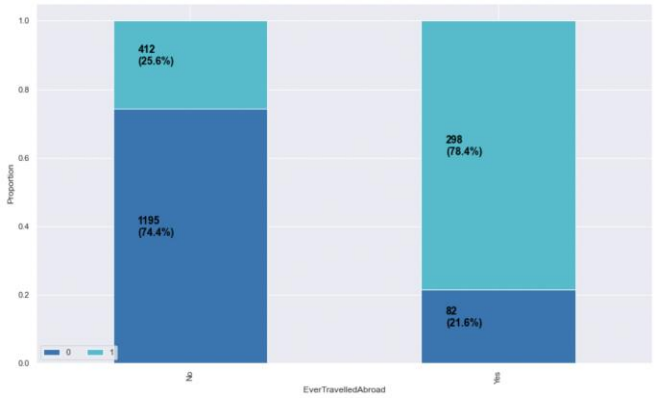




EverTravelledAbroad

We modified the variable with 1 representing customers who have travelled to foreign countries, and 0 representing customers who have not travelled to foreign countries. Out of 1987 observations, 380 (19.1%) have travelled abroad, 298 (78.4%) of which purchased travel insurance, and 1607 (80.9%) have not travelled abroad, 412 (25.6%) of which purchased travel insurance. Although the proportion of people who have travelled abroad is small, our stacked bar chart indicates that customers are a lot more likely to purchase travel insurance if they have travelled abroad.





IV. Variable Selection

We would like to see whether all 8 predictors are necessary and useful in predicting our target variable. To accomplish this, we used the LassoCV function within the sklearn.linear_model class. Trying to minimize the cost function, Lasso regression will automatically select those features that are useful, discarding the useless or redundant features. Based on the result, EverTravelledAbroad and AnnualIncome are the most important predictors, followed by FamilyMembers, Age and FrequentFlyer. GradueteOrNot, Employment.Type and ChronicDiseases seem to be insignificant in predicting whether someone will purchase travel

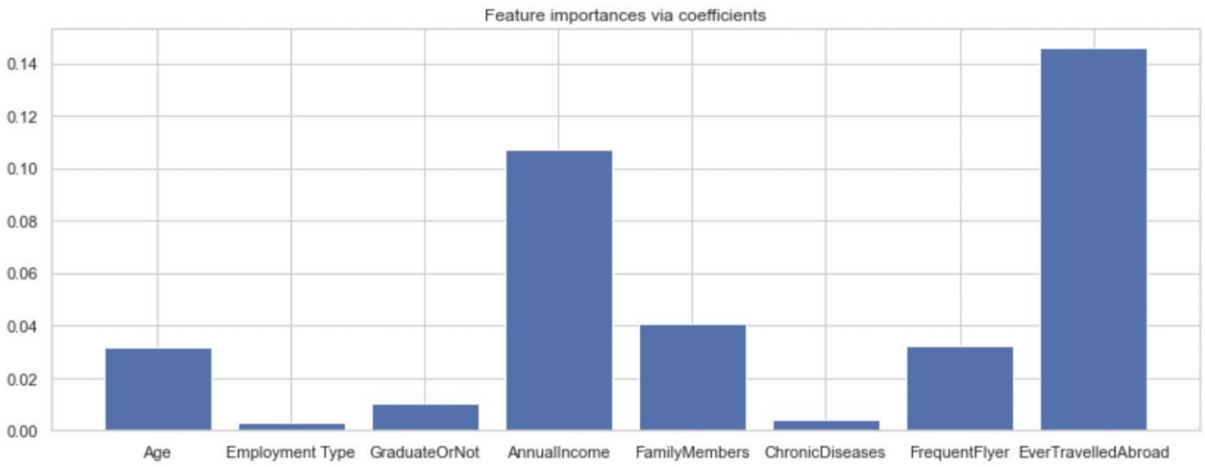
insurance. In the next step, we will first try to fit our models with all 8 predictors, then the top 5 important predictors and top 3 important predictors and see how the results change.

V. Models

Split Train and Test Data

Before fitting any models, we split the dataset into two parts, where 20% of the dataset will be used as a test set, and the rest will be used as a training set. Dividing the dataset is necessary to validate our model and ensure that it will generalize to unseen data.

Logistic Regression



Logistic regression predicts probability of binary event occurrence by using the logit function. As one of the generalized linear models, it is relatively easy to interpret and has relatively high predicting power. We first performed logistic regression using the function from the sklearn package. Using all 8 predictors, the prediction accuracy on the test set is **76.63%**. Using the top five and top three predictors, the accuracy on the test set is **77.14%** and **76.38%**, respectively. Since including more predictors does not lead to a significant increase in prediction accuracy, if we were to use the logistic regression model, we would prefer the one with three predictors: EverTravelledAbroad, AnnualIncome and FamilyMembers.

Neural Network

We applied gradient descent to find the optimal weights by iteratively minimizing the cost function. Setting the learning rate at 0.1 and the iteration at 100, the resulting prediction accuracy on the test set with all predictors is **77.89%**. With five predictors, the accuracy remained at **77.89%**, confirming that GraduateOrNot, Employment.Type and ChronicDiseases have little impact. With three predictors, the test accuracy is **77.14%**. Again, we prefer the model with three predictors, since the accuracy does not suffer significantly, while the interpretability of the model increases.

KNN

The K-nearest neighbor is a non-parametric learning algorithm, which makes it useful for real world data. It calculates the distance between new data points and existing data, then predicts it to be the same as the majority of K-nearest data points. Since training is not required for making the prediction, the implementation of KNN is relatively fast and easy. Using the KNeighborsClassifier function from the sklearn package with k=2, the test accuracy using eight, five and three predictors is **78.89%**, **77.64%**, and **77.14%**, respectively. In order to enhance the performance of our model, we try to find the optimal k value by plotting the k values from 1 to 19 and corresponding accuracy measures. The optimal k value that results in the highest accuracy is k=4. Then the improved test accuracy using eight, five, and three predictors is **80.65%**, **81.91%**, and **79.90%**, respectively.

Decision Tree

Decision tree uses recursive binary splitting to split the data into a finite set of non-overlapping regions. All observations in the same buckets will have the same predicted value. Since it does not depend on any probability distribution assumptions, it is less computationally intensive compared to some other algorithms like neural networks. Decision trees are even easier to interpret than regression models because the binary splits minic human decision-making processes and it can also be displayed visually. Furthermore, it automatically performs feature selection by not including the insignificant predictors. We first created decision trees using the DecisionTreeClassifier function from sklearn package. Using eight, five and three predictors, the results are all very large trees with test accuracy of **76.63%**, **76.13%**, and **78.39%**, respectively.

Pruned Decision Tree

Since the initial tree is complicated and likely to overfit, pruning is necessary to reduce the size of a tree and remove less valuable splits. This process reduces overfitting and can lead to a simpler, more interpretable tree. With a cost complexity parameter of 0.01, the pruned trees from eight and five predictors now have only three splits using three predictors and four terminal nodes with a test accuracy of **82.16%**. Surprisingly, the three predictors selected in the order of importance are AnnualIncome, Age, and FamilyMembers. The predictor EverTravelledAbroad is not chosen, despite that we initially thought it was significant. The pruned tree from three predictors now has only two splits and three terminal nodes with a test accuracy of **80.4**%. The two automatically chosen predictors are AnnualIncome and FamilyMembers.

Bagging

Bagging combines the results of a set of decision trees fitted to a different bootstrapped sample of the training data, then using the average to make a final prediction. It reduces overfitting and variance of the base tree, thus leading to higher prediction accuracy. However, the disadvantage is that it loses the interpretability of decision trees and is computationally intensive. The resulting test accuracy of bagging models with eight, five and three predictors is **79.15%**, **80.65%**, and **77.64%**, respectively.

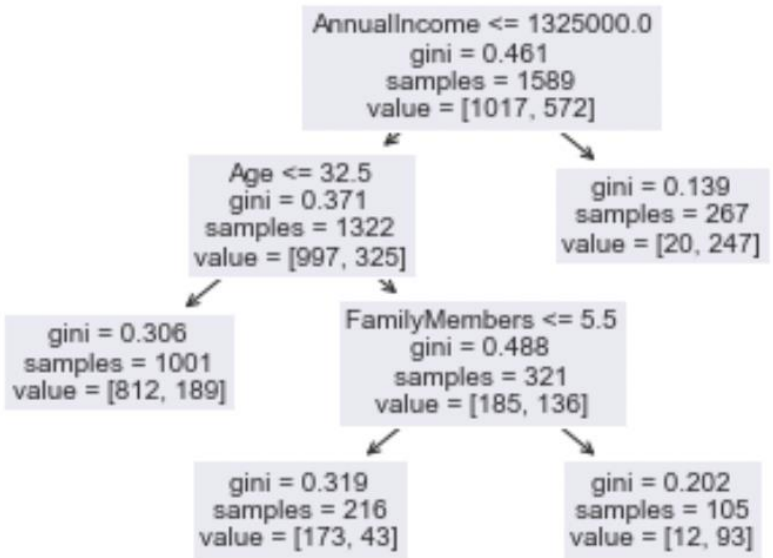
VI. Model Performance and Selection

In the table below, we compared the performance of 17 models that we fitted using three measures: accuracy, sensitivity and specificity. Sensitivity is the frequency of correctly predicting an event of interest when that event occurs. For instance, a sensitivity of 51.45% means that we would be able to identify 51.45% of the customers who intend to purchase travel insurance. Specificity is the frequency of correctly predicting a non-event when there is indeed no occurrence of that event. For instance, a specificity of 90% means that we would be able to identify 90% of the customers who do not intend to purchase travel insurance. Accuracy is a weighted measure of sensitivity and specificity, identifying the total percentage of correct predictions made. In the context of our business problem, we care more about sensitivity since we want to identify potential customers to sell our products to. However, one consistent pattern from all our models is that the sensitivity is relatively low, ranging from 50.00% to 65.22%, while our specificity is relatively high, ranging from 84.23% to 95.77%. This is due to the imbalance of our data, where the proportion of people who purchased travel insurance is only 35.7%. Ideally, we want a proportion close to 50%, which will increase the sensitivity and accuracy of our models. We will suggest some ways to accomplish this in the future research section.

Models	Accuracy	Sensitivity	Specificity
Logistic Regression (8 predictors)	76.63%	51.45%	90.00%
Logistic Regression (5 predictors)	77.14%	51.45%	90.77%
Logistic Regression (3 predictors)	76.38%	50.00%	90.38%
Neural Network (8 predictors)	77.89%	53.62%	90.77%
Neural Network (5 predictors)	77.89%	54.34%	90.38%
Neural Network (3 predictors)	77.14%	52.17%	90.38%
KNN (k=4, 8 predictors)	80.65%	60.14%	85.38%
KNN (k=4, 5 predictors)	81.91%	61.59%	85.38%
KNN (k=4, 3 predictors)	79.90%	57.25%	85.77%
Decision Tree (8 predictors)	76.63%	65.22%	84.23%
Decision Tree (5 predictors)	76.13%	57.97%	85.77%
Decision Tree (3 predictors)	78.39%	53.62%	91.54%
Pruned Decision Tree (3 predictors)	**82.16%**	**57.97%**	**95.00%**

Pruned Decision Tree (2 predictors)	80.40%	51.45%	95.77%
Bagging (8 predictors)	79.15%	64.49%	86.92%
Bagging (5 predictors)	80.65%	64.49%	89.23%
Bagging (3 predictors)	77.64%	55.80%	89.23%

Disregarding the issue with low sensitivity, we want to select a final model with a relatively high accuracy and interpretability. Our final recommendation is the pruned decision trees with three predictors: AnnualIncome, Age, and FamilyMembers. Not only does it have the highest test accuracy of 82.16%, it is also the simplest model to explain to a general audience who is not familiar with statistical concepts, such as company management or the marketing team. As shown in the graph, if the annual income of a customer is above 1,325,000 Rupees, we predict that they will purchase travel insurance; if the annual income of the customer is less than or equal to 1,325,000 Rupees, we then look at the age of the customer. If the customer's age is less than or equal to 32.5, we predict that they will not purchase the travel insurance; otherwise, we look at the number of members in the customer's family. If it is five or below, we predict that they will not purchase travel insurance; otherwise, we predict that they will purchase travel insurance.



VII. Conclusion

Our objective in this project is to build a model using some or all the given variables to predict whether a customer is likely to purchase travel insurance. We first checked that our dataset is complete and usable, and modified the format of some variables so they can be easily used. We

randomly split the dataset with 80% of the data in the training set, and 20% of the data in the test set. We then went through the variable selection process using data visualization and some built-in functions. Several models were considered and trained, each fitted three times, first with all the available predictors, then with five and three selected predictors. After comparing the prediction accuracy on the testing dataset, we decided that the decision tree with pruning is the best model. Not only does it achieve the highest accuracy of 82.16% on the test set, it is also the most interpretable model, since it only uses three predictors and is easy to present visually.

VIII. Future Research

Sampling Techniques

As previously discussed, the imbalance data resulted in low sensitivity. One solution is to use sample techniques such as oversampling, undersampling, or a combination of both to make the data more balanced. Oversampling duplicates observations in the minority class (people who purchased travel insurance) until the proportions are roughly equal, while undersampling removes some observations from the majority class (people who did not purchase travel insurance) until the proportions are roughly equal. Since undersample results in loss of information, our intuition is that oversampling will likely improve our model performance. This question can also be solved by including a larger dataset that is representative of the customer base.

Other Target of Interest

In addition to predicting whether someone will purchase travel insurance, we would like to suggest the company look into another cost-related target. For instance, for those who did purchase travel insurance, what were their premium amounts or coverage amount? Knowing the most applicable predictors that contribute to a high coverage insurance plan would help the company prioritize prospective clients.

Other Potential Predictors

Here are some other factors that might be potential predictors for whether someone will purchase travel insurance, or the amount of coverage that they will purchase.

- Distance - numeric variable of total distance travelled.
- Duration - numeric variable of number of nights spent on the trip.
- Reason - categorical variable of main reasons for the trip; sample categories might include vacation, visit, business, etc.
- Timing - categorical variable of the timing of travel, such as in months or quarters.
- Mode - categorical variable of main mode of transportation; sample categories might include car, train, plan, cruise, etc.
- Region - categorical variable of destination regions, with an intuition that claim costs are correlated with the cost of living, safety, environment and other factors that are associated with a region.

Other Possible Models

- Elastic Net Model
- Boosting
- Support Vector Machines

IX. Appendix & Reference

[1]. Travel Insurance Prediction.html
[2]. Travel Insurance Prediction.ppt
[3]. TravelInsurancePrediction.csv
[4]. https://www.kaggle.com/tejashvi14/travel-insurance-prediction-data