DSC 550

Marc Riley

Travel Insurance

Every year nearly 4 billion dollars is spent on travel insurance in the US (Newsweek 2018). There is great potential for travel insurance companies to make even more money due to the fact that only 35% of all travelers buy insurance for their trip.

If we could potentially find travelers that would be interested in purchasing insurance at a more successful rate millions of dollars could be made. Using machine learning and data mining we will attempt to group travelers together that are more likely to by insurance off of many different factors.

The Data

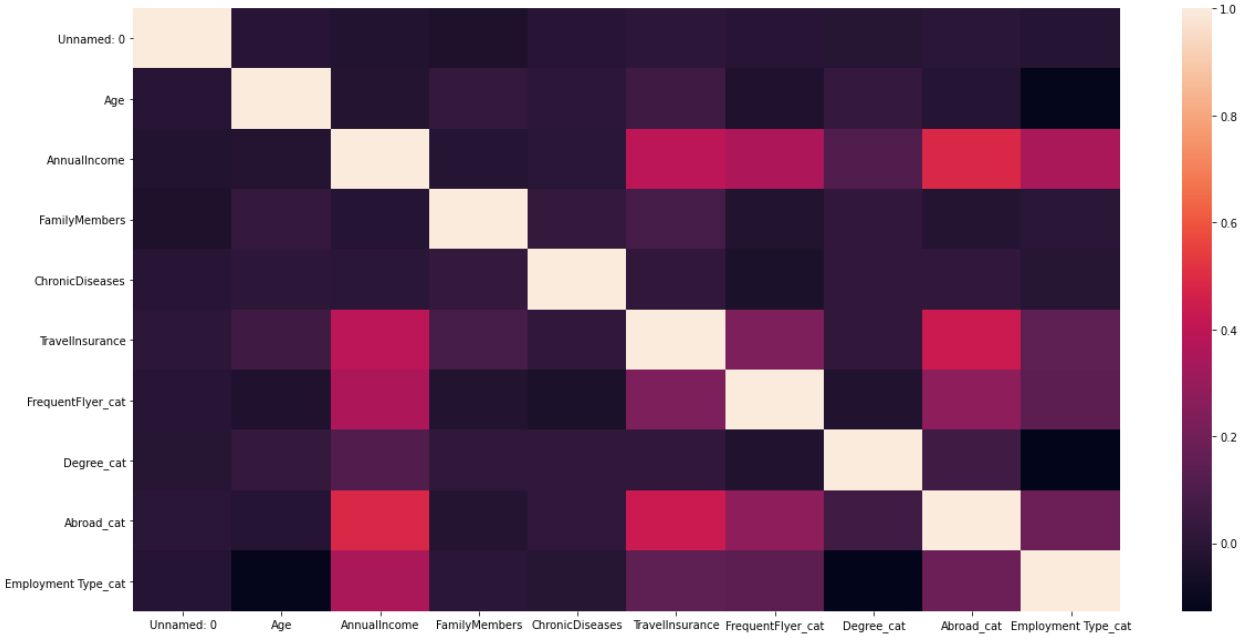
The data set being used contains 1987 cases of individuals that regularly buy or don't buy travel insurance. Of these nearly 2 thousand records 709 of them have bought insurance and 1276 have not. Some of the information that is provided on these individuals include

- Age
- Income
- Education Level
- If they have a chronic disease
- Frequent Flyer Status
- If they Travel Abroad
- Number of Family Members

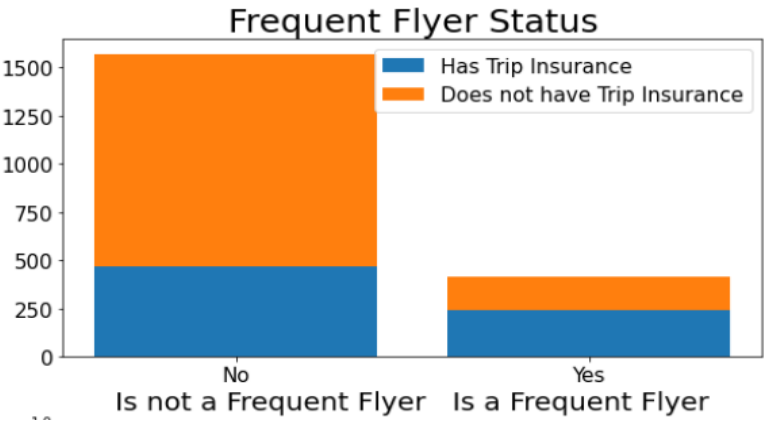
The data is complete and no missing values needed to be adjusted for.

Graphical Analysis

Graphical analysis is an important step in any project. It gives us valuable insights that might not be clear by looking at the raw data. The first step that was take was to view the correlation between some of the variables. This heat map shows which variables correlate. Before I was to get this the data had to be transformed to make the categorical features into binary ones.

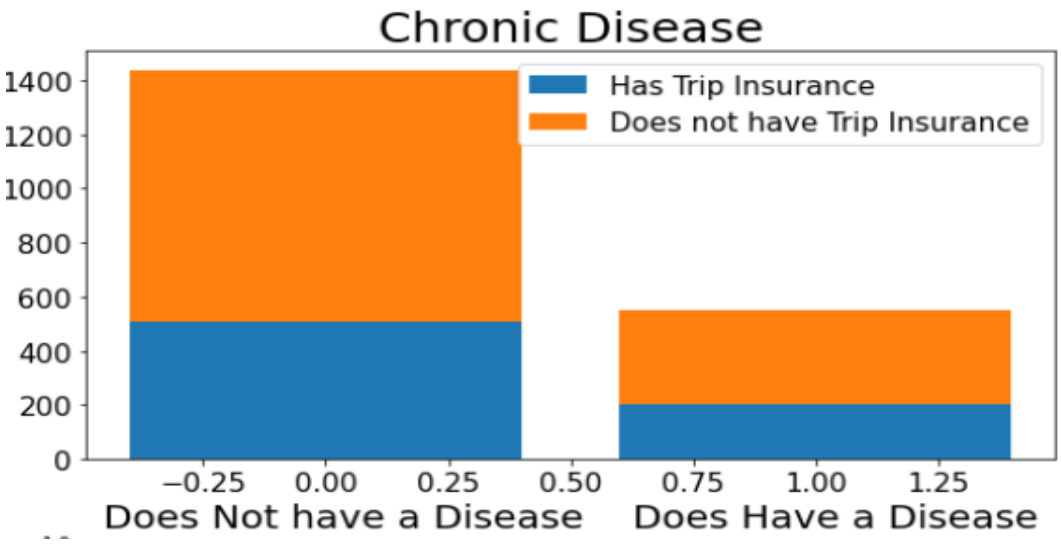


Using the correlation chart we can see that the frequent flyer category correlates to whether someone buys travel insurance. A better way to see this is with a bar chart. As we can see below more than half of all frequent flyers purchase travel insurance.



Feature Selection

Also known as variable selection is the process of getting a subset of the data that only includes relevant variables or predictors. This is done to improve our algorithms efficacy and accuracy. I took a univariate approach. I removed features such as the degree variable and chronic diseases since they do not seem to play a role in weather or not travel insurance is purchased. The bar graph below shoes people with chronic diseases chose travel insurance abut the same rate as those without.



Model Selection/Evaluation

The model picked for this project is a Decision Tree. This model works by breaking down our dataset into smaller subsets giving us a result of a tree with multiple decision and leaf nodes. This model works well with our data because it can handle both categorical and numerical data. After training our model and tweaking the parameters we can make it 83% accurate. Giving us the ability to accurately predict when someone will purchase travel insurance based off of our variables.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.98      0.89       383
           1       0.95      0.58      0.72       214

    accuracy                           0.84       597
   macro avg       0.88      0.78      0.80       597
weighted avg       0.86      0.84      0.83       597

Accuracy:  0.8391959798994975
```

Conclusion

This project had its challenges from converting categorical features to finding and implementing a model. Overall I believe my decision tree produced decent results. It is very accurate when predicting when someone will not buy insurance. I tried to implement a random forest model but was unsuccessful. The current application for this project could help travel insurance companies identify new customers.