Data Science Capstone Project

Capstone Project Assessment

Travel Insurance Dataset

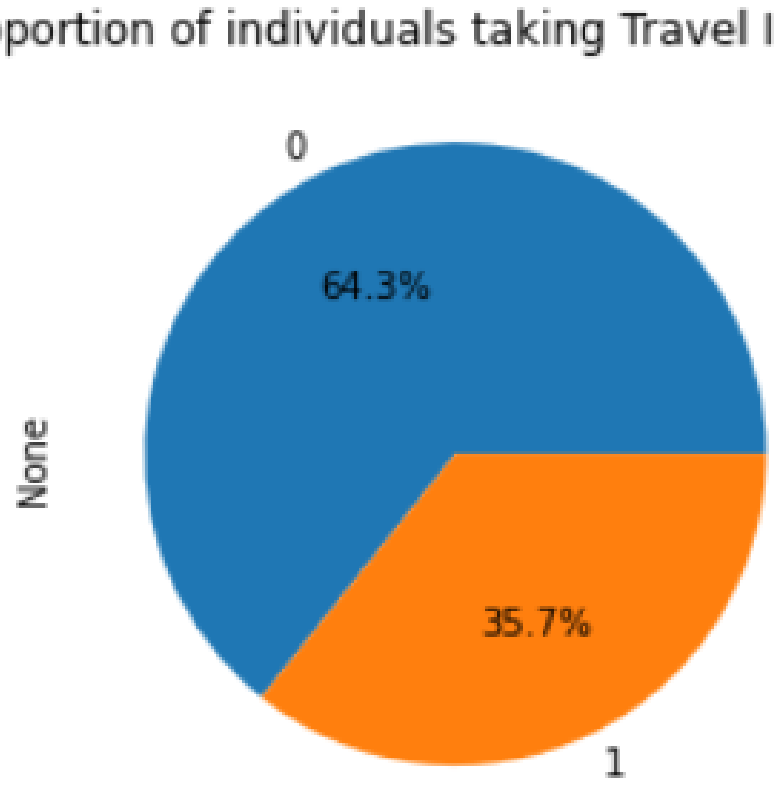
Problem Statement

- Travel Insurance is an important purchase for consumers when it comes to travel planning. Travel Insurance expenditure is heavily linked to consumer requirements. Based on personal requirements of their age, health status, family demographics, travel frequencies, consumers select the options for Travel Insurance products available on the market to suit their needs.
- Analyzing current datasets based on factors such as age groups, employment status, Annual Income amounts, education levels, health statuses, family backgrounds helps to provide information on their spending patterns.
- Models can be generated to predict consumer spending trends. Such trends help provide Travel Insurance providers design better products with suitable customizable options. This can better cater to consumer needs thereby improving service efficiency and quality levels.

Dataset

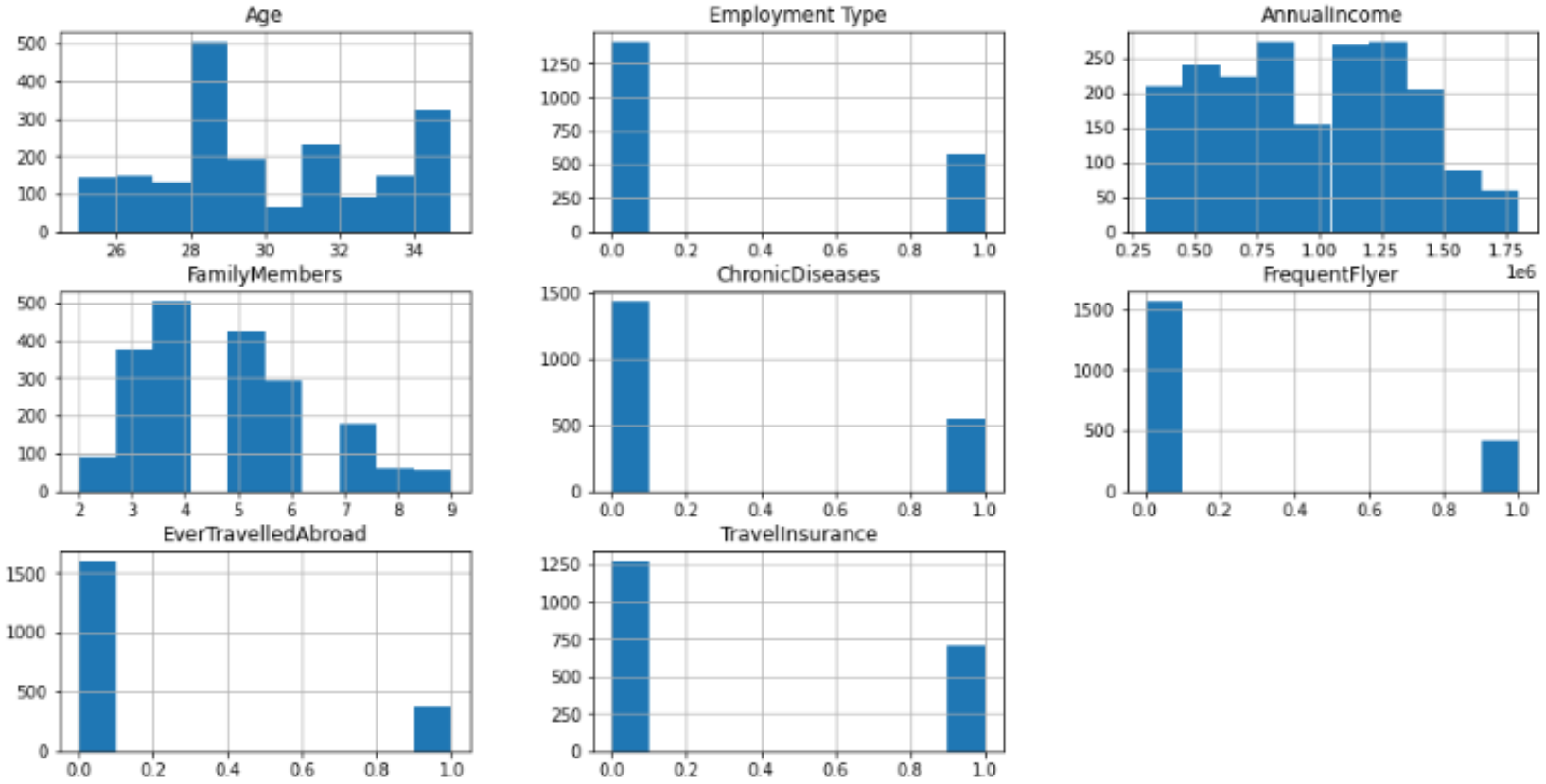
```
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unnamed: 0            1987 non-null   int64
 1   Age                   1987 non-null   int64
 2   Employment Type       1987 non-null   object
 3   GraduateOrNot         1987 non-null   object
 4   AnnualIncome          1987 non-null   int64
 5   FamilyMembers         1987 non-null   int64
 6   ChronicDiseases       1987 non-null   int64
 7   FrequentFlyer         1987 non-null   object
 8   EverTravelledAbroad   1987 non-null   object
 9   TravelInsurance       1987 non-null   int64
dtypes: int64(6), object(4)
```

Proportion of individuals taking Travel Insurance



Out of a data set contain 1987 cases,
Only about 35,7% buy travel insurance

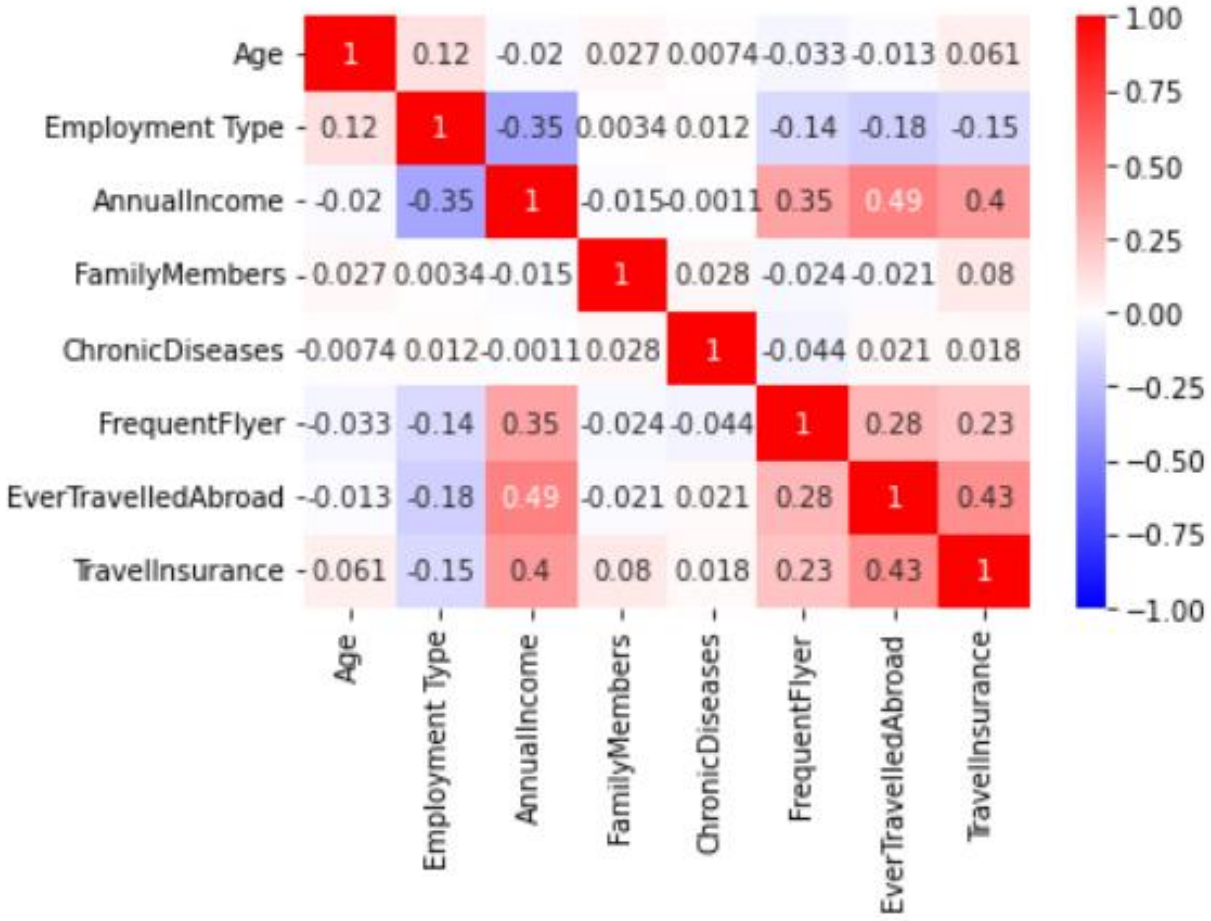
Histogram plot



Age and Annual Income are Continuous

Employment Type, Chronic Diseases Frequent Flyer and
Ever travelled Abroad are categorical

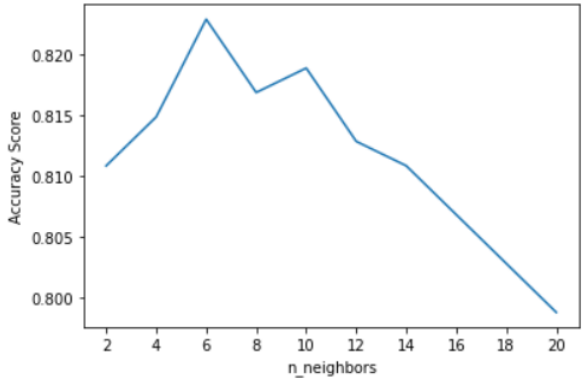
Correlation Chart



Strongest Correlation is observed
Ever Travelled Abroad-Annual Income

Models applied

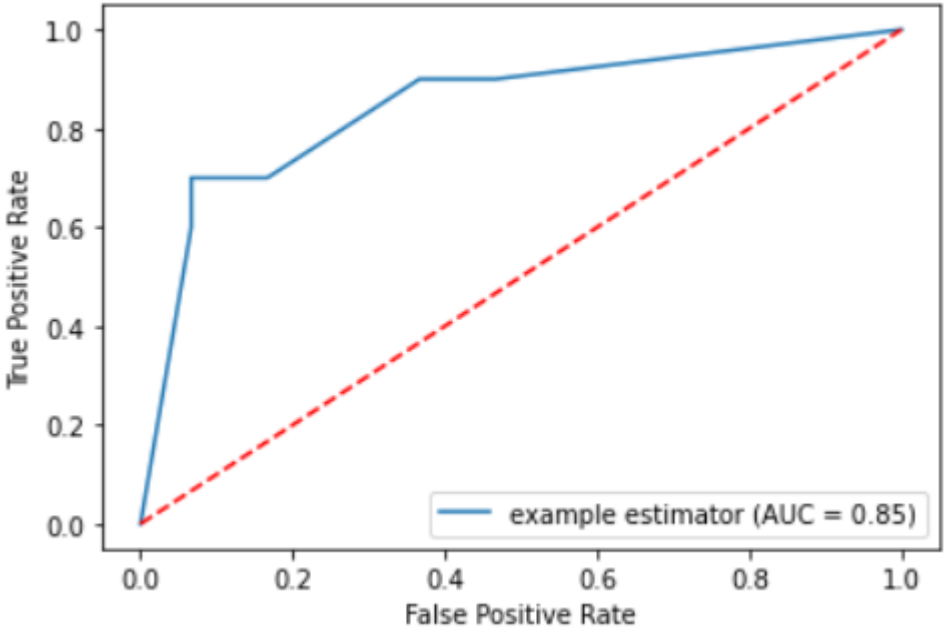
- As this is a categorical dataset, the following classification models are applied
  - Logistic Regression -> Accuracy Score: 0.659
  - K Nearest Neighbors -> Accuracy Score: 0.823



  - Decision Tree -> Accuracy Score: 0.845

Decision Tree Improvement

- Decision Tree produces best performance score out of the 3 models with 0.845 accuracy score
- Further improvement was performed using Grid Search
- Grid Search improvement -> Accuracy Score: 0.846



Conclusion

- Applied Logistic regression, KNN, Decision Tree to determine which is best model.
- Decision tree model provided best accuracy score of 0.8451.
- Improvement of performance using GridSearch optimize Decision Tree model and improved accuracy score to 0.8466.
- For further model improvement, we can join this dataframe to other data, to create more complete data set. More modelling can be performed with improved accuracy obtained.