Prediction of Travel Insurance Purchase

Group E

PoAn Chen, Susan Wang,
Wenyu Wu, Jing Zhang

ACTUPS 5841

Professor： Yubo Wang

12/9/2021

Agenda

- Project Introduction
- Data Overview
- Variable Selection
- Models
- Conclusion
- Future research
- Q&A

Project Introduction

Travel insurance reimburses policy holders for:

- lost baggage
- flight delays
- medical problems

Real dataset and business problem from a tour & travels company in India

Data Overview

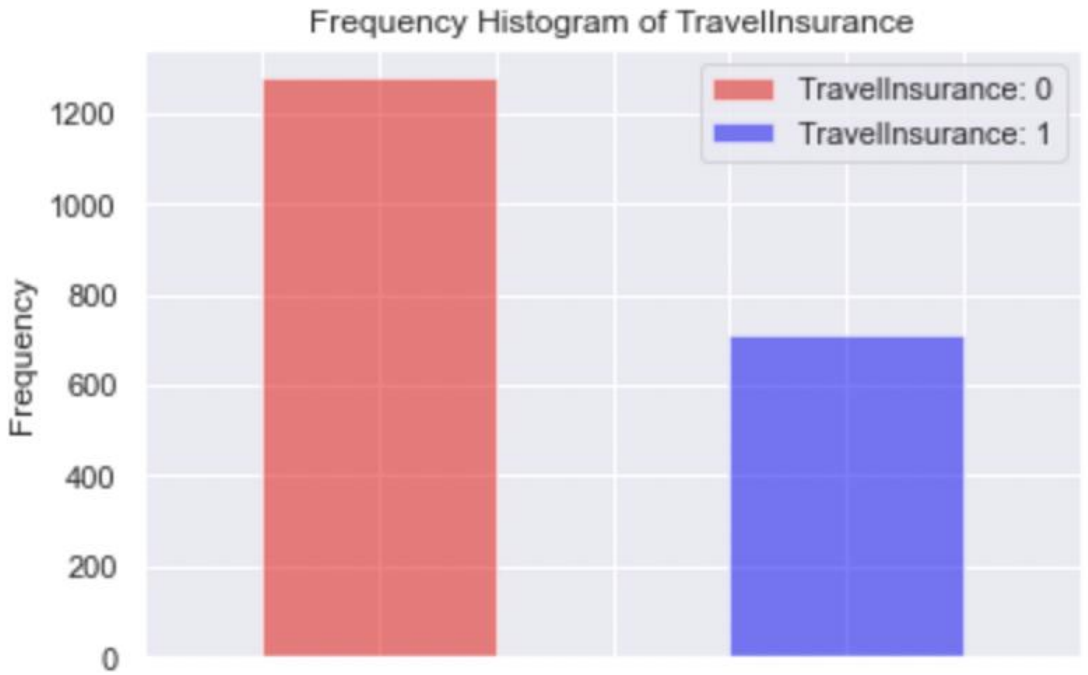
Total: 1987 observations

Target	Travel Insurance	1 if the customer purchased travel insurance package in 2019, and 0 otherwise
Predictors	Age	Age of the customer
	Employment Type	The sector in which customer is employed
	Graduate Or Not	Whether the customer is college graduate or not
	Annual Income	Annual income of the customer in Indian Rupees (rounded to nearest 50 thousand Rupees)
	Family Members	Number of members in customer's family
	Chronic Disease	1 if the customer suffers from major disease or conditions like diabetes, high blood pressures, asthma, etc, and 0 otherwise
	Frequent Flyer	Whether the customer booked at least 4 flight tickets in the past two years (2017-2019)
	Ever Travelled Abroad	Whether the customer has travelled to a foreign country

Data Visualization

Target: Travel Insurance (TI)

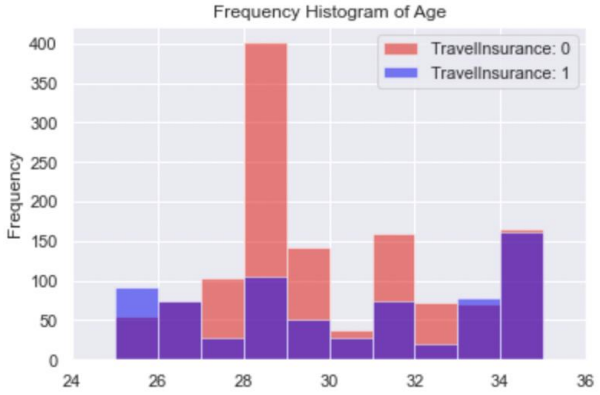
- Out of 1987 observations
- 710 (35.7%) purchased TI
- 1277 (64.3%) did not purchase TI

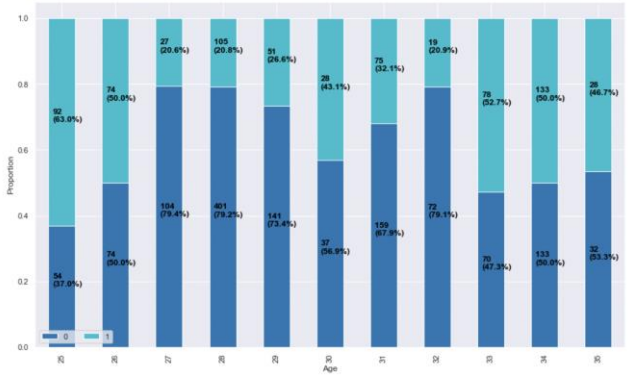


Data Visualization

Age

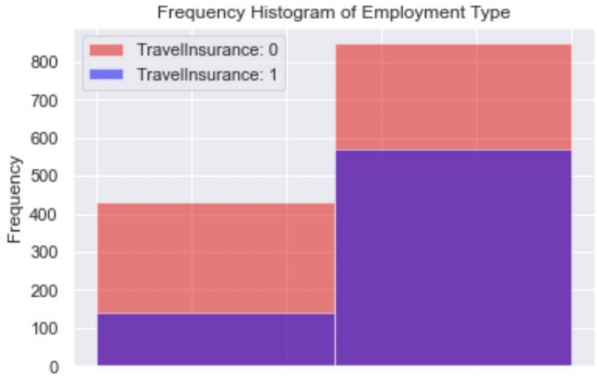
TravelInsurance	0	1
Age		
25	54	92
26	74	74
27	104	27
28	401	105
29	141	51
30	37	28
31	159	75
32	72	19
33	70	78
34	133	133
35	32	28

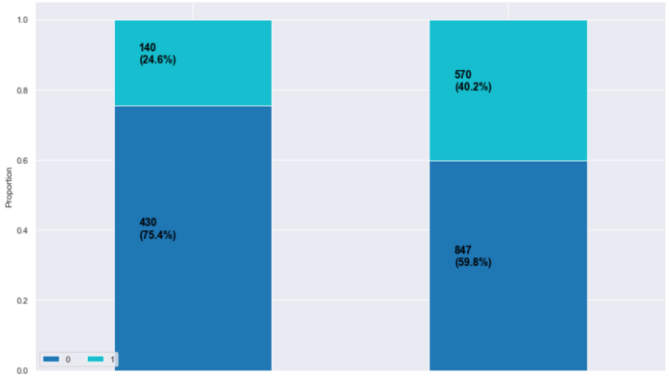




Employment Type

TravelInsurance	0	1
Employment Type		
Government Sector	430	140
Private Sector/Self Employed	847	570

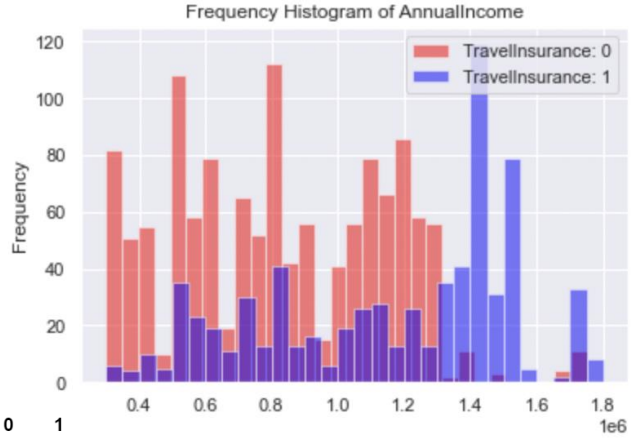


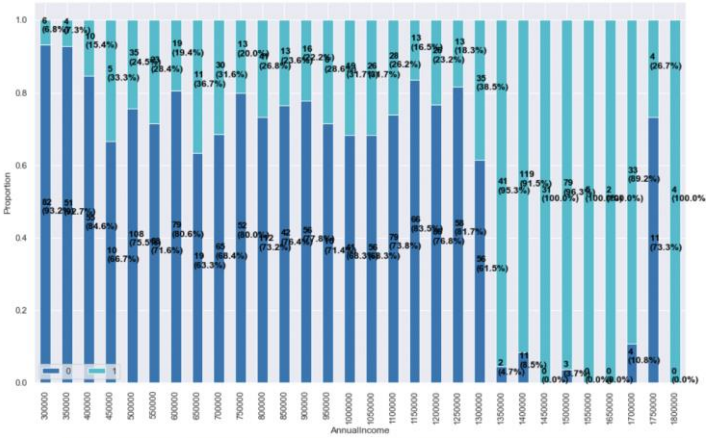


Data Visualization

Annual Income (in Rupees)

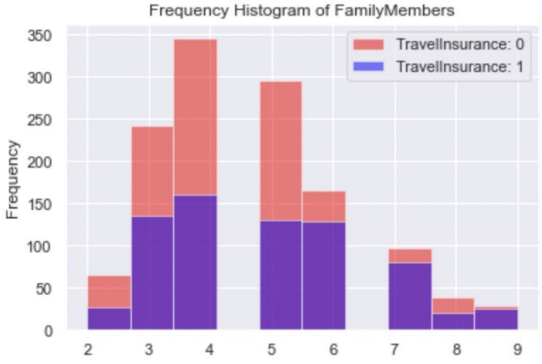
- Range: [300,000, 1,800,000]

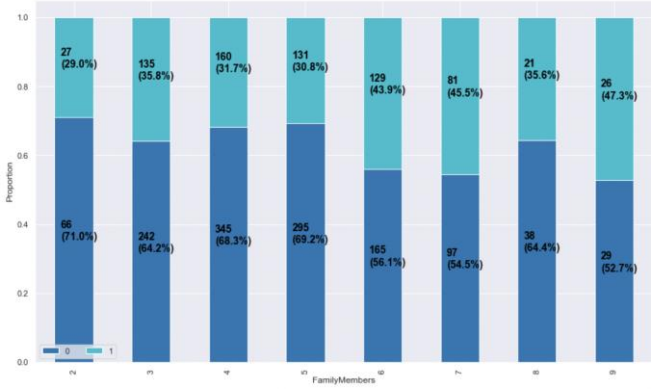




Family Members

TravelInsurance	0	1
FamilyMembers		
2	66	27
3	242	135
4	345	160
5	295	131
6	165	129
7	97	81
8	38	21
9	29	26





Agenda

- Project Introduction
- Data Overview
- Variable Selection
- Models
- Conclusion
- Future research
- Q&A

Project Introduction

Travel insurance reimburses policy holders for:

- lost baggage
- flight delays
- medical problems

Real dataset and business problem from a tour & travels company in India

Data Overview

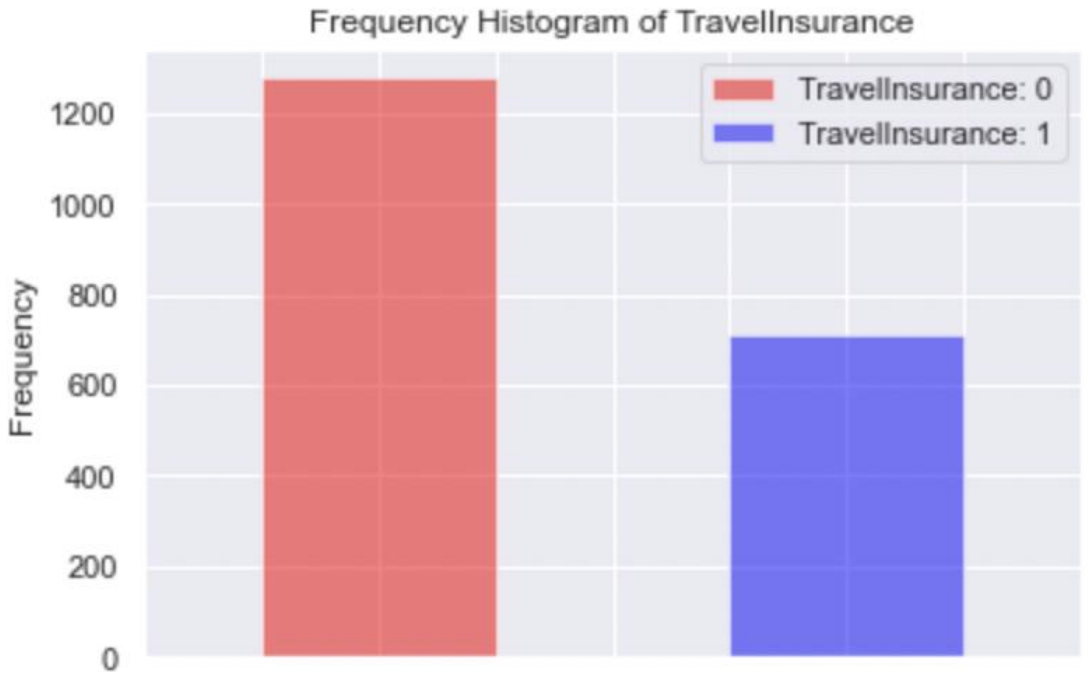
Total: 1987 observations

Target	Travel Insurance	1 if the customer purchased travel insurance package in 2019, and 0 otherwise
Predictors	Age	Age of the customer
	Employment Type	The sector in which customer is employed
	Graduate Or Not	Whether the customer is college graduate or not
	Annual Income	Annual income of the customer in Indian Rupees (rounded to nearest 50 thousand Rupees)
	Family Members	Number of members in customer's family
	Chronic Disease	1 if the customer suffers from major disease or conditions like diabetes, high blood pressures, asthma, etc, and 0 otherwise
	Frequent Flyer	Whether the customer booked at least 4 flight tickets in the past two years (2017-2019)
	Ever Travelled Abroad	Whether the customer has travelled to a foreign country

Data Visualization

Target: Travel Insurance (TI)

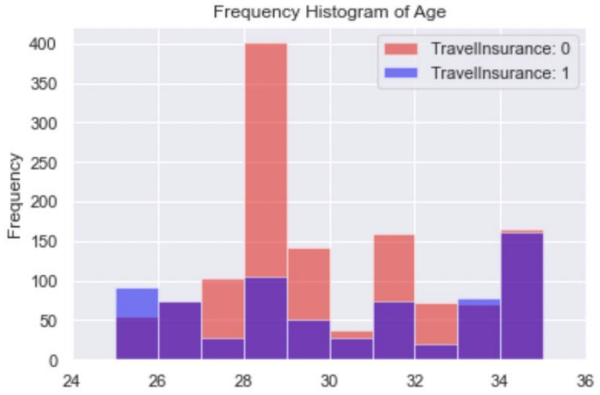
- Out of 1987 observations
- 710 (35.7%) purchased TI
- 1277 (64.3%) did not purchase TI

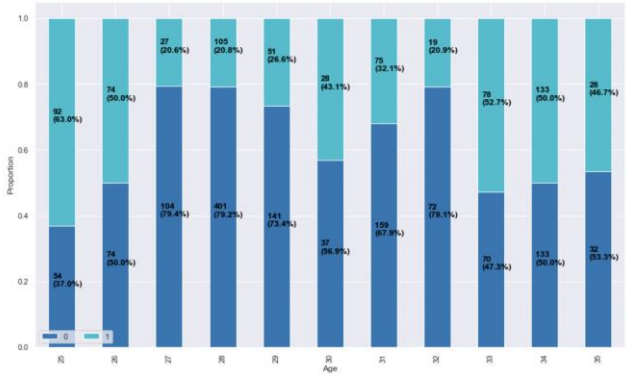


Data Visualization

Age

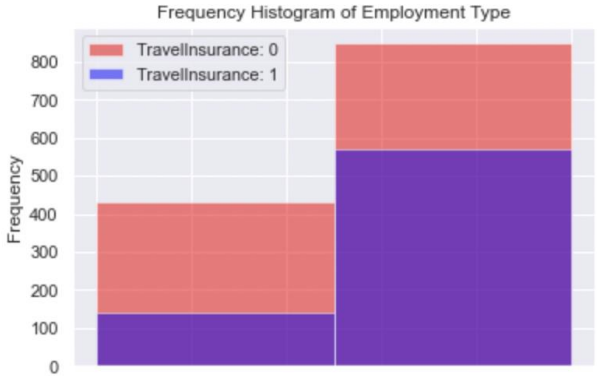
TravelInsurance	0	1
Age		
25	54	92
26	74	74
27	104	27
28	401	105
29	141	51
30	37	28
31	159	75
32	72	19
33	70	78
34	133	133
35	32	28

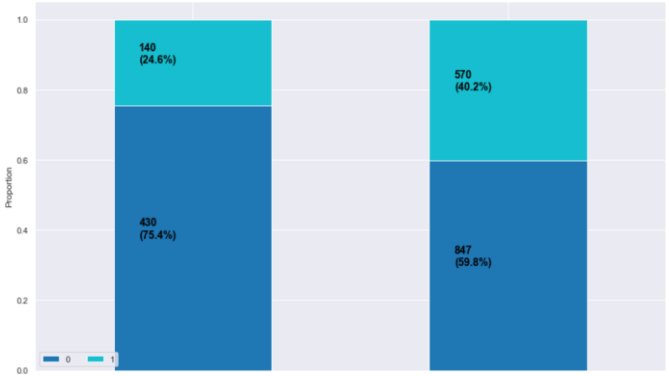




Employment Type

TravelInsurance	0	1
Employment Type		
Government Sector	430	140
Private Sector/Self Employed	847	570

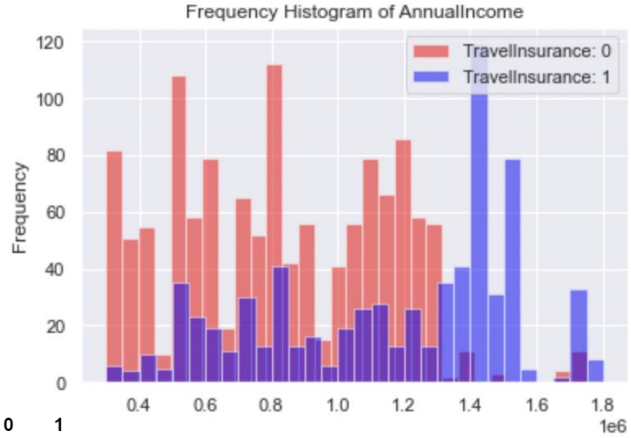


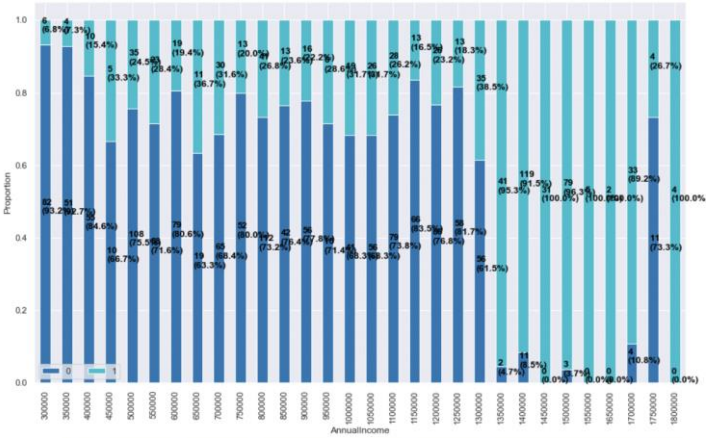


Data Visualization

Annual Income (in Rupees)

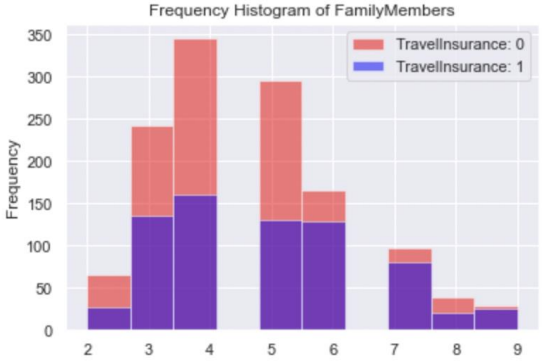
- Range: [300,000, 1,800,000]

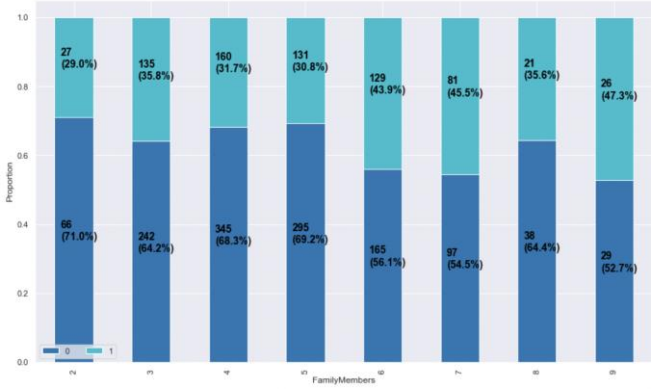




Family Members

TravelInsurance	0	1
FamilyMembers		
2	66	27
3	242	135
4	345	160
5	295	131
6	165	129
7	97	81
8	38	21
9	29	26

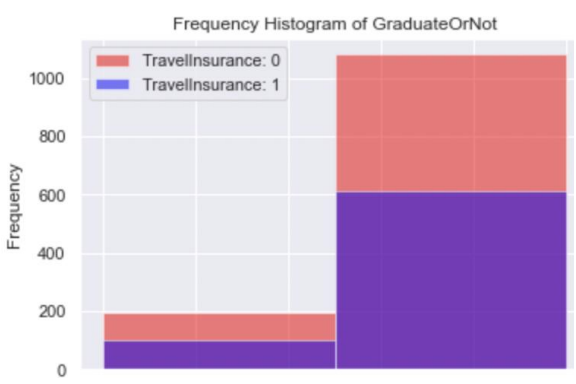


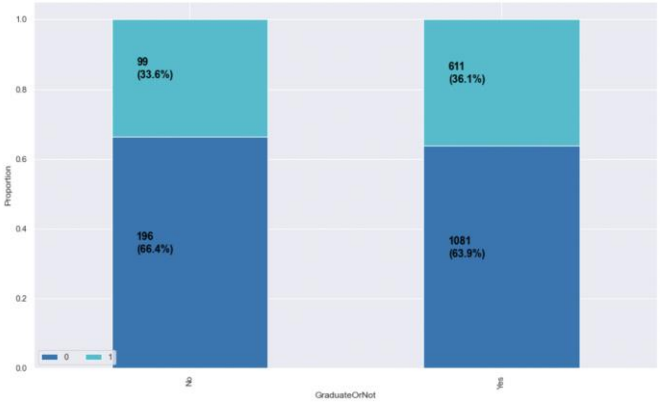


Data Visualization

Graduate Or Not

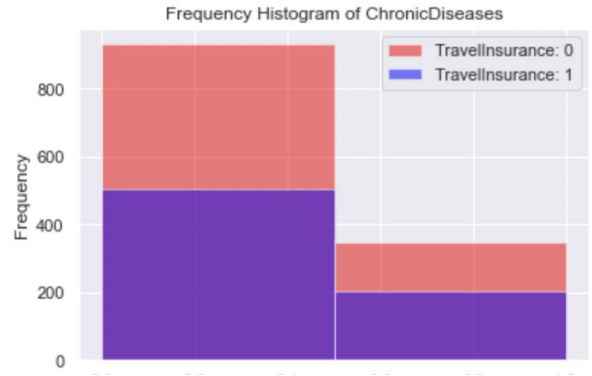
TravelInsurance	0	1
GraduateOrNot		
No	196	99
Yes	1081	611

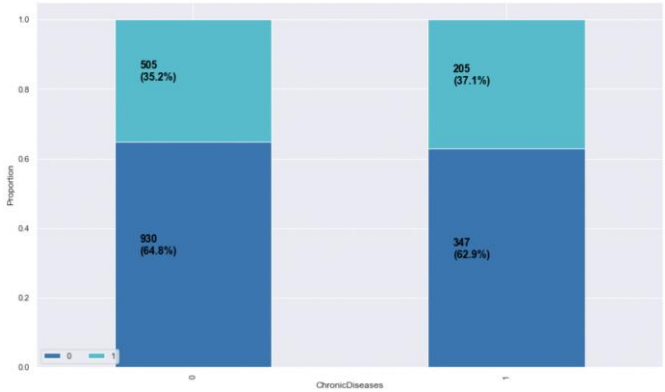




Chronic Disease

TravelInsurance	0	1
ChronicDiseases		
0	930	505
1	347	205

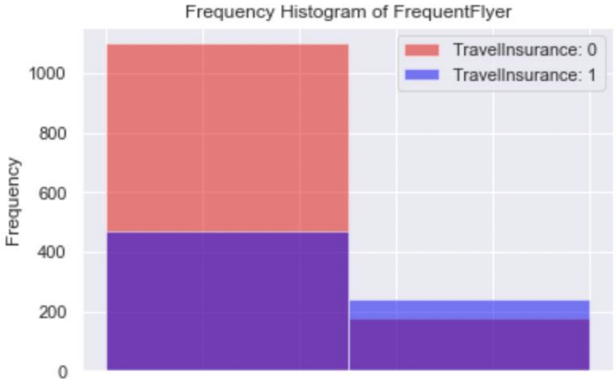


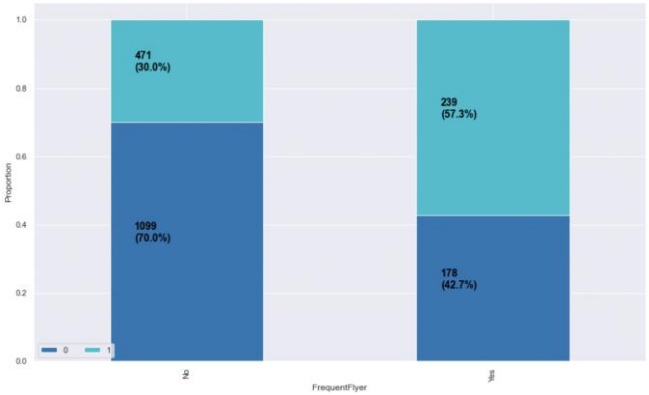


Data Visualization

Frequent Flyer

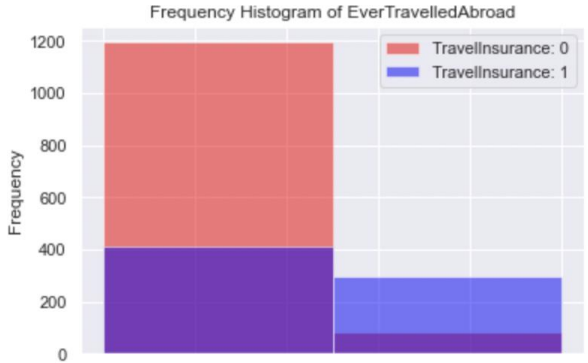
TravelInsurance	0	1
FrequentFlyer		
No	1099	471
Yes	178	239

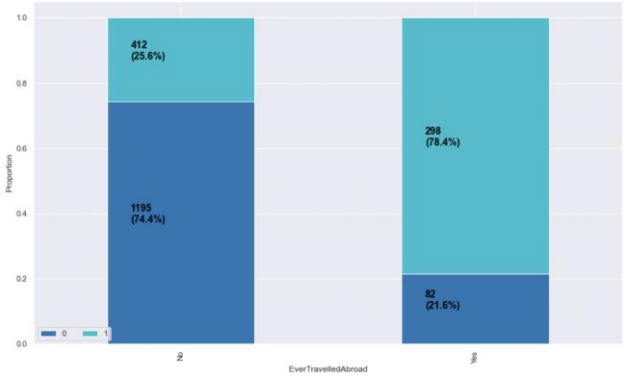




Ever Travelled Abroad

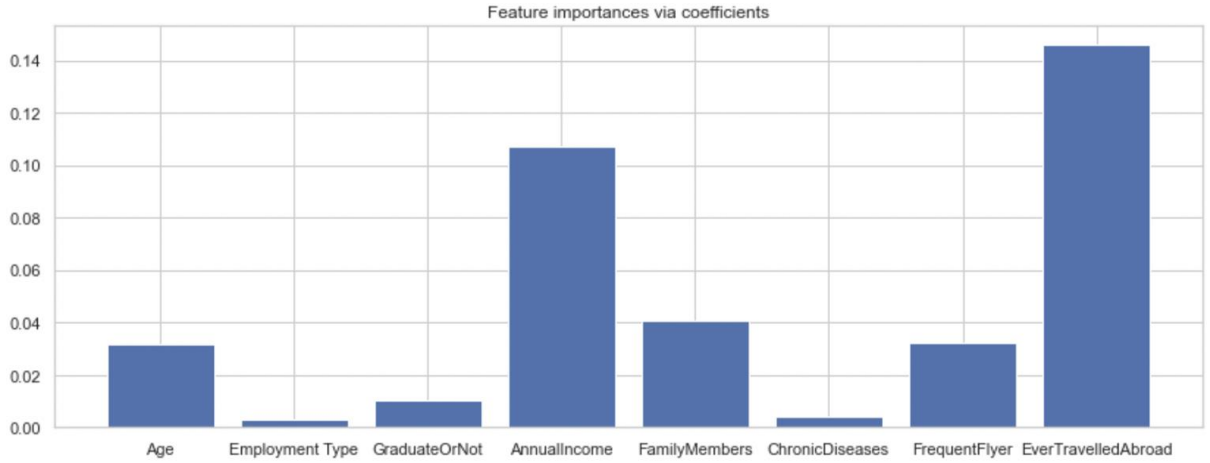
TravelInsurance	0	1
EverTravelledAbroad		
No	1195	412
Yes	82	298





Feature Selection

- Ever Travelled Abroad and Annual Income are the most important predictors, followed by Family Members, Age and Frequent Flyer.
- Graduate Or Not, Employment Type and Chronic Diseases seem to be insignificant.



Models

Logistic Regression

All predictors

Test Accuracy : 76.63%

(EverTravelledAbroad,
AnnualIncome,
FamilyMembers,
Age,
FrequentFlyer,
Graduate Or Not,
Employment type,
Chronic Diseases)

5 predictors

Test Accuracy : 77.14%

(EverTravelledAbroad,
AnnualIncome,
FamilyMembers,
Age,
FrequentFlyer.)

3 predictors

Test Accuracy : 76.38%

(EverTravelledAbroad,
AnnualIncome,
FamilyMembers)

Agenda

- Project Introduction
- Data Overview
- Variable Selection
- Models
- Conclusion
- Future research
- Q&A

Project Introduction

Travel insurance reimburses policy holders for:

- lost baggage
- flight delays
- medical problems

Real dataset and business problem from a tour & travels company in India

Data Overview

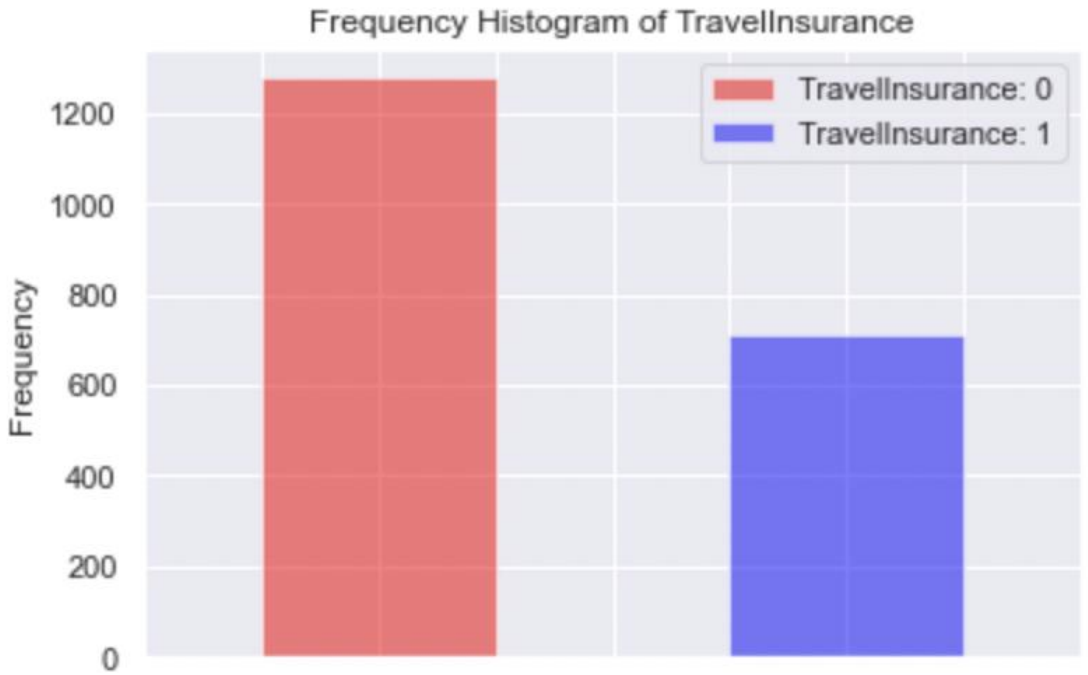
Total: 1987 observations

Target	Travel Insurance	1 if the customer purchased travel insurance package in 2019, and 0 otherwise
Predictors	Age	Age of the customer
	Employment Type	The sector in which customer is employed
	Graduate Or Not	Whether the customer is college graduate or not
	Annual Income	Annual income of the customer in Indian Rupees (rounded to nearest 50 thousand Rupees)
	Family Members	Number of members in customer's family
	Chronic Disease	1 if the customer suffers from major disease or conditions like diabetes, high blood pressures, asthma, etc, and 0 otherwise
	Frequent Flyer	Whether the customer booked at least 4 flight tickets in the past two years (2017-2019)
	Ever Travelled Abroad	Whether the customer has travelled to a foreign country

Data Visualization

Target: Travel Insurance (TI)

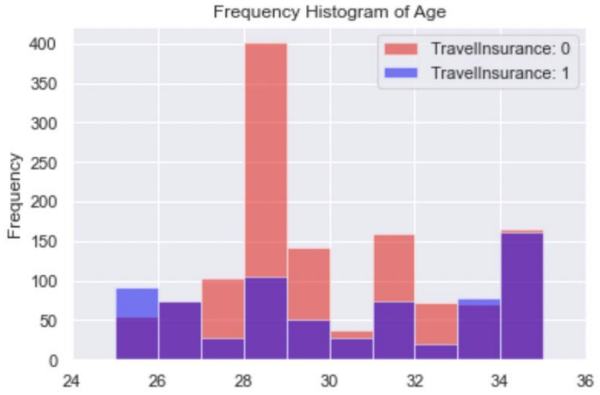
- Out of 1987 observations
- 710 (35.7%) purchased TI
- 1277 (64.3%) did not purchase TI

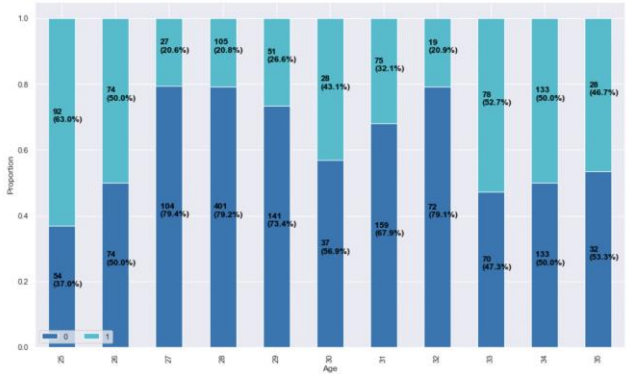


Data Visualization

Age

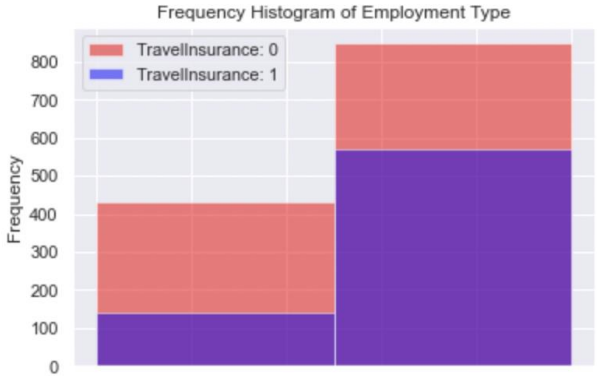
TravelInsurance	0	1
Age		
25	54	92
26	74	74
27	104	27
28	401	105
29	141	51
30	37	28
31	159	75
32	72	19
33	70	78
34	133	133
35	32	28

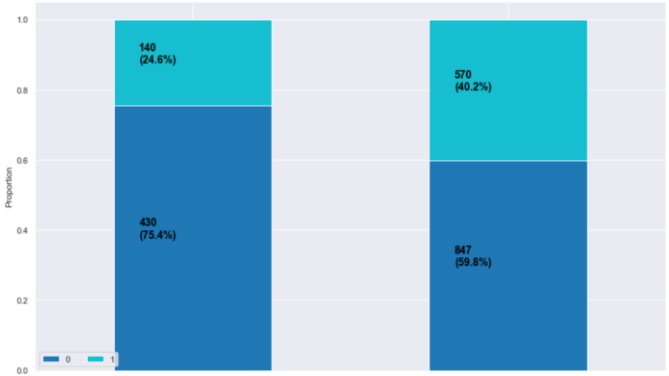




Employment Type

TravelInsurance	0	1
Employment Type		
Government Sector	430	140
Private Sector/Self Employed	847	570

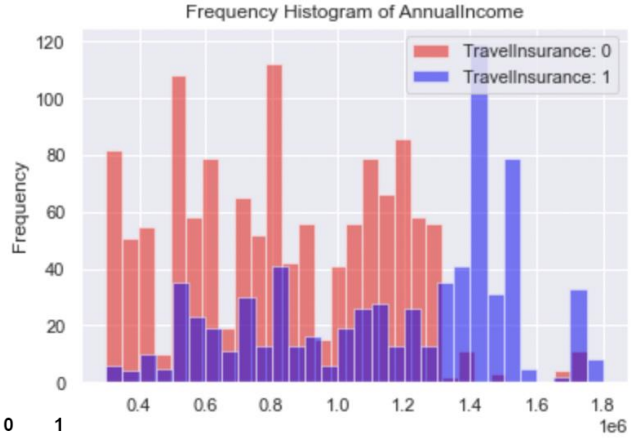


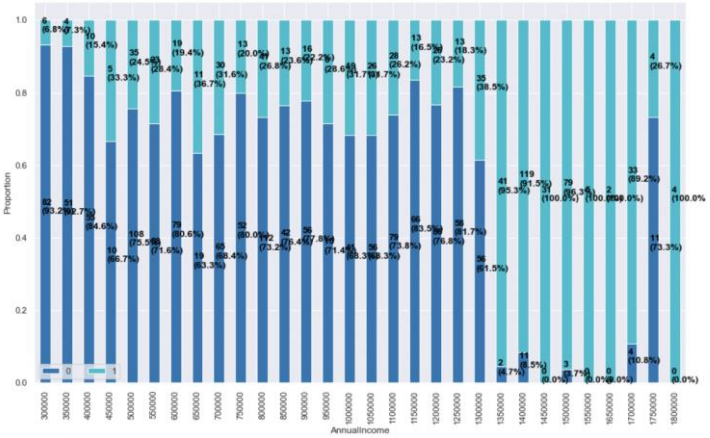


Data Visualization

Annual Income (in Rupees)

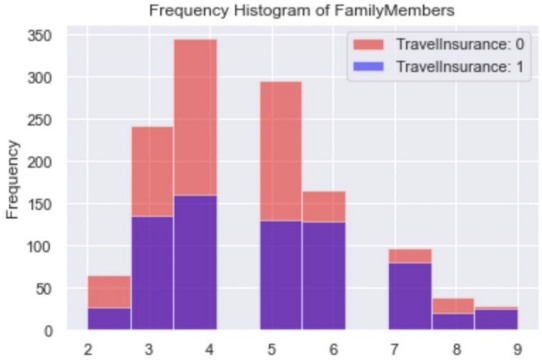
- Range: [300,000, 1,800,000]

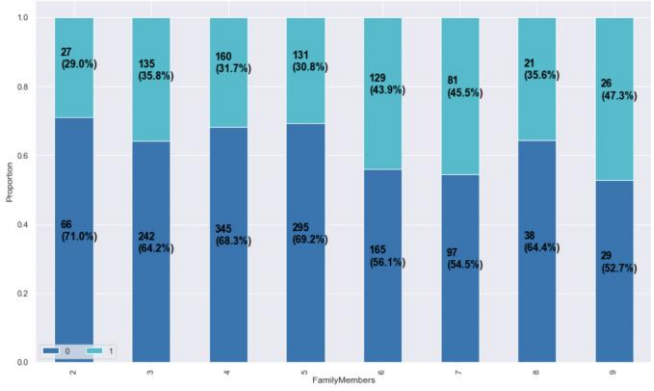




Family Members

TravelInsurance	0	1
FamilyMembers		
2	66	27
3	242	135
4	345	160
5	295	131
6	165	129
7	97	81
8	38	21
9	29	26

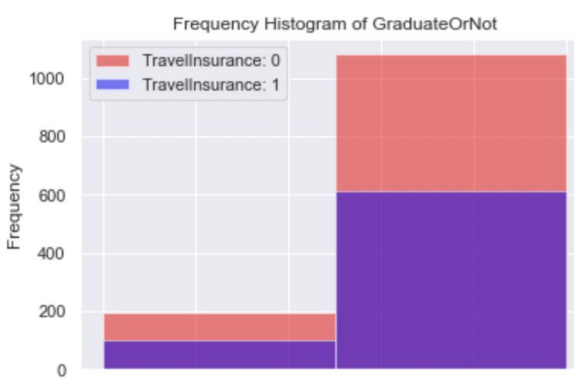


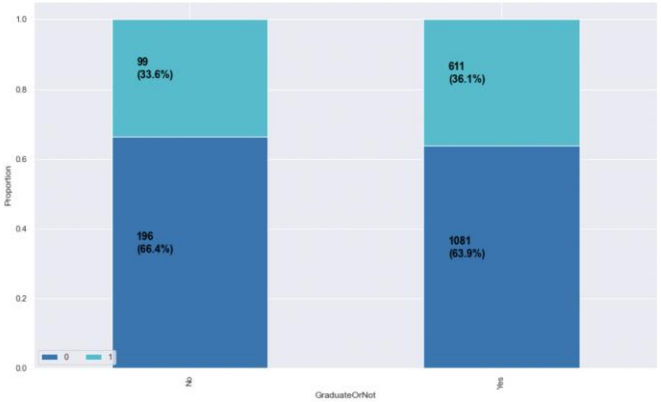


Data Visualization

Graduate Or Not

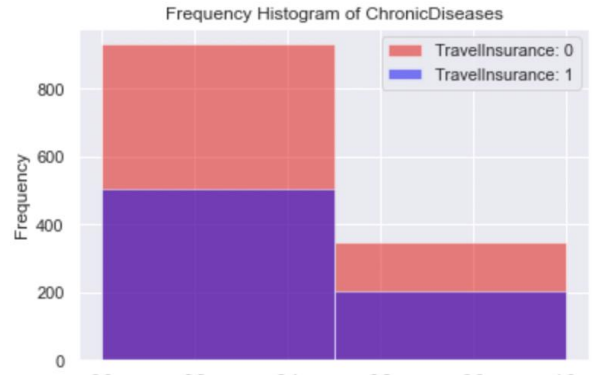
TravelInsurance	0	1
GraduateOrNot		
No	196	99
Yes	1081	611

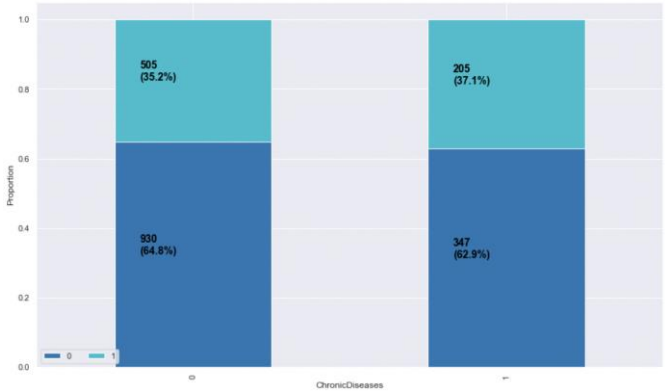




Chronic Disease

TravelInsurance	0	1
ChronicDiseases		
0	930	505
1	347	205

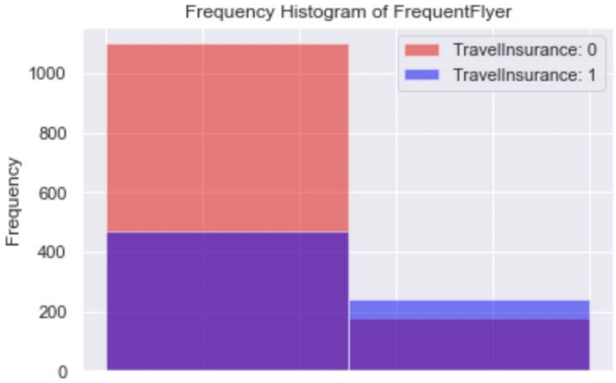


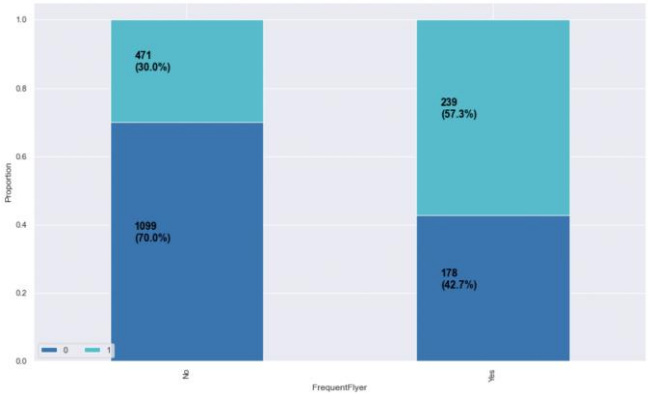


Data Visualization

Frequent Flyer

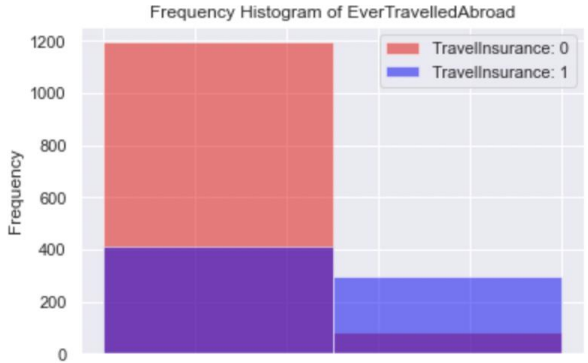
TravelInsurance	0	1
FrequentFlyer		
No	1099	471
Yes	178	239

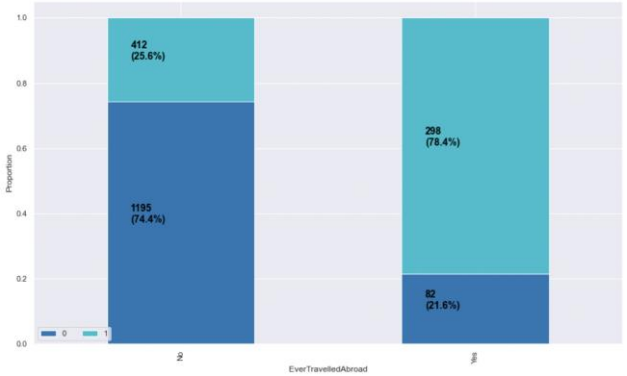




Ever Travelled Abroad

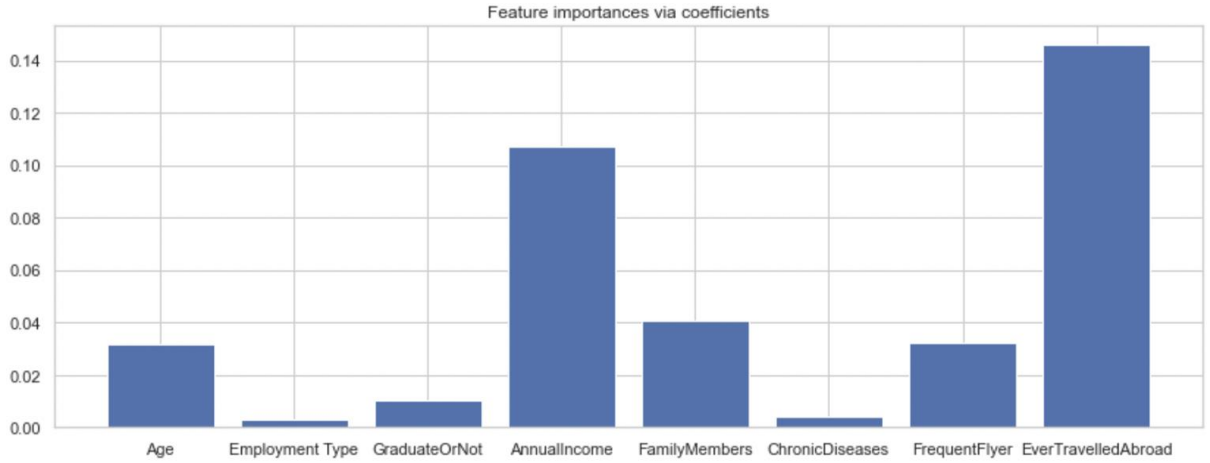
TravelInsurance	0	1
EverTravelledAbroad		
No	1195	412
Yes	82	298





Feature Selection

- Ever Travelled Abroad and Annual Income are the most important predictors, followed by Family Members, Age and Frequent Flyer.
- Graduate Or Not, Employment Type and Chronic Diseases seem to be insignificant.



Models

Logistic Regression

All predictors

Test Accuracy : 76.63%

(EverTravelledAbroad,
AnnualIncome,
FamilyMembers,
Age,
FrequentFlyer,
Graduate Or Not,
Employment type,
Chronic Diseases)

5 predictors

Test Accuracy : 77.14%

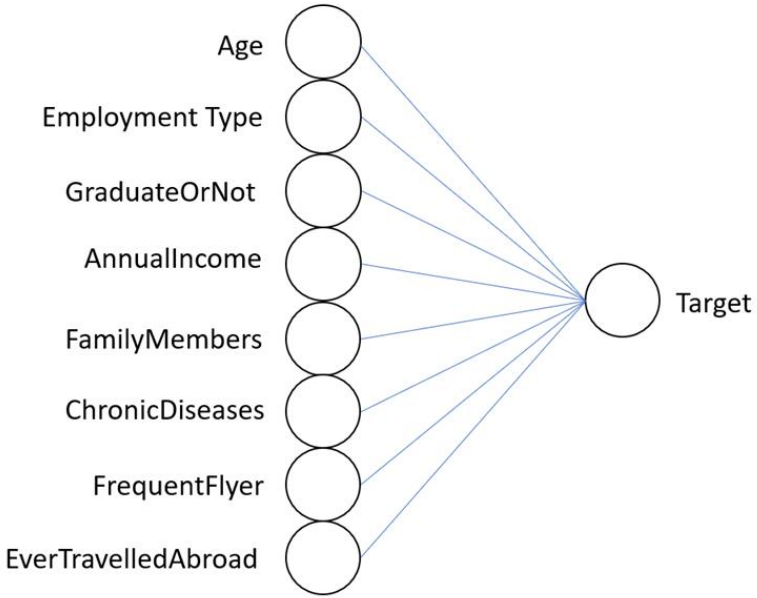
(EverTravelledAbroad,
AnnualIncome,
FamilyMembers,
Age,
FrequentFlyer.)

3 predictors

Test Accuracy : 76.38%

(EverTravelledAbroad,
AnnualIncome,
FamilyMembers)

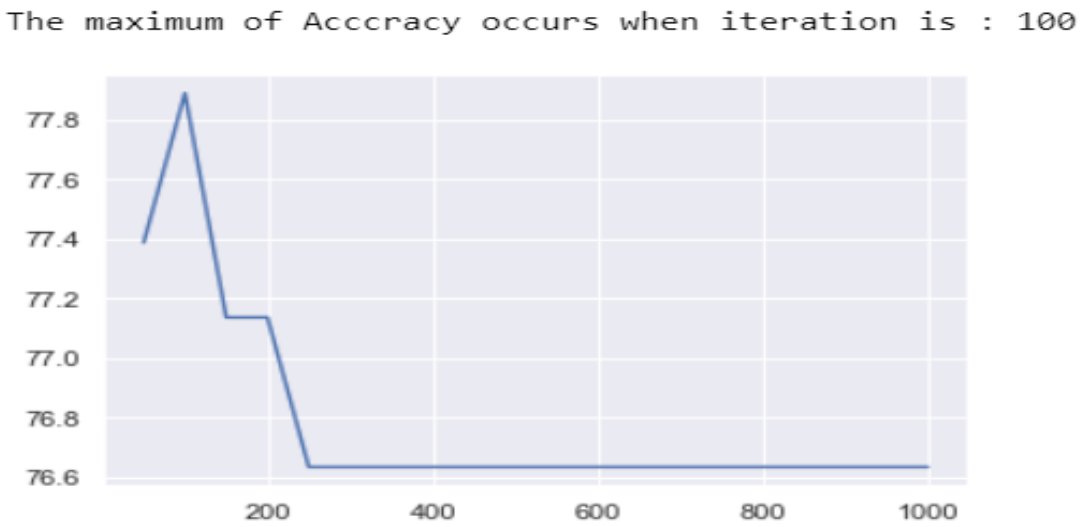
Logistic Regression-Neural network model



Learning rate:10%
Initial weight:0.01
Initial bias:0.00
Sigmoid activation function

Logistic Regression-Neural network

Neural network model runs 100 iteration can get the highest accuracy

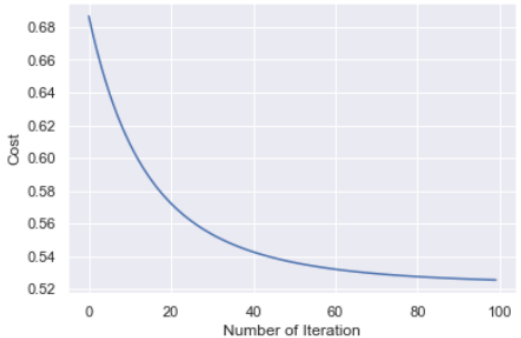


Logistic Regression-Neural network

All predictors

Test Accuracy : 77.89%

```
iteration: 100
cost: 0.525468061646591
```

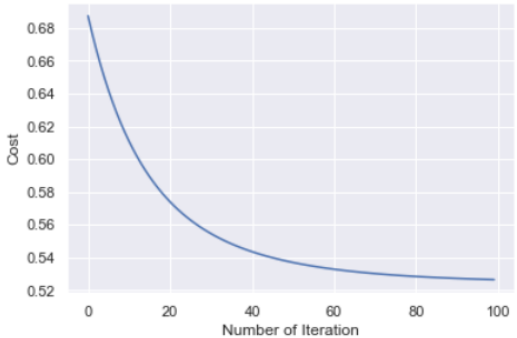



```
Manuel Test Accuracy: 77.89%
```

5 predictors

Test Accuracy : 77.89%

```
iteration: 100
cost: 0.5263271311902352
```

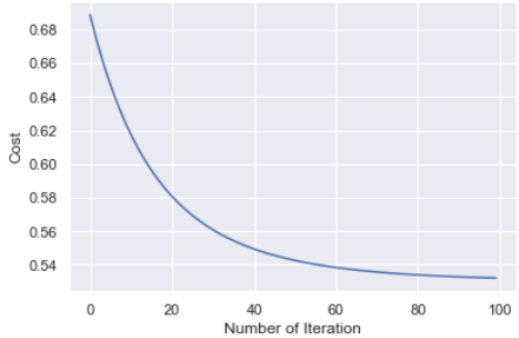



```
Manuel Test Accuracy: 77.89%
```

3 predictors

Test Accuracy : 77.14%

```
iteration: 100
cost: 0.5321283157357334
```




```
Manuel Test Accuracy: 77.14%
```

KNN (K=2)

All predictors

2 NN Score: 78.89%

5 predictors

2 NN Score: 77.64%

3 predictors

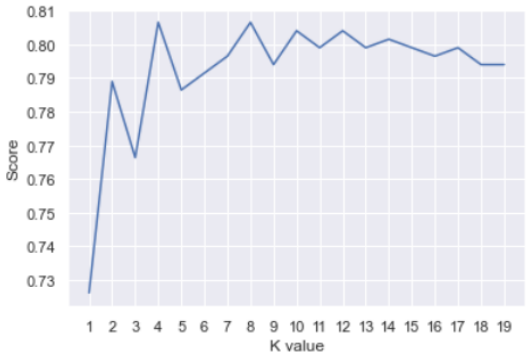
2 NN Score: 77.14%

KNN-finding best K

All predictors

1.K=4

2.KNN Score is 80.65%

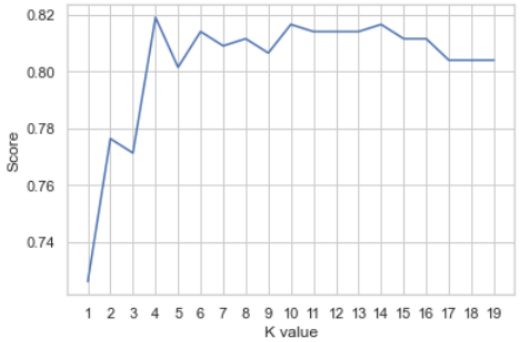



```
Maximum KNN Score is 80.65%
The Maximun KNN is  4
```

5 predictors

1.K=4

2.KNN Score is 81.91%

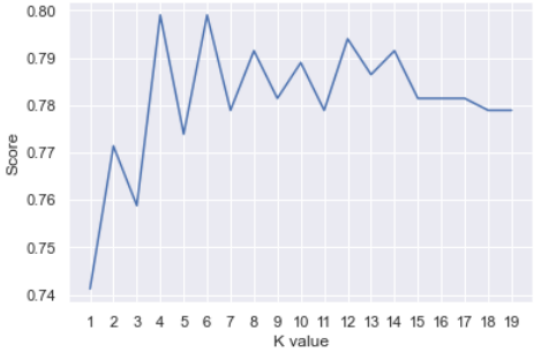



```
Maximum KNN Score is 81.91%
The Maximun KNN is  4
```

3 predictors

1.K=4

2.KNN Score is 79.90%



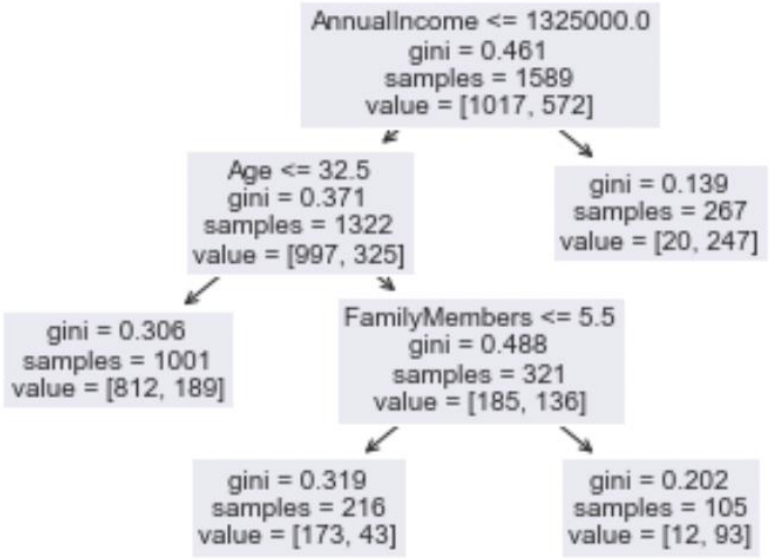

```
Maximum KNN Score is 79.90%
The Maximun KNN is  4
```

Decision Tree

- Recursive binary splitting process that splits data into a finite set of non-overlapping regions.
- Pro: easy to interpret and display; no probability distribution assumption
- Con: vulnerable to overfitting.
- All predictors: 76.63%, five predictors: 76.13%, three predictors: 78.39%.

Pruned Decision Tree

- Pruning is necessary to reduce the size of a tree and remove less valuable splits.
- Pro: reduces overfitting and can lead to a simpler, more interpretable tree; automatically performs variable selection.
- All predictors and five predictors: pruned tree has three predictors with 82.16% accuracy.
- Three predictors: pruned tree has two predictors with 80.4% accuracy.



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.
- Pro: reduces overfitting and variance of the base tree, leading to higher prediction accuracy.
- Con: loses the interpretability of decision trees and is computationally intensive.
- All predictors: 79.15%, five predictors: 80.65%, three predictors: 77.64%.

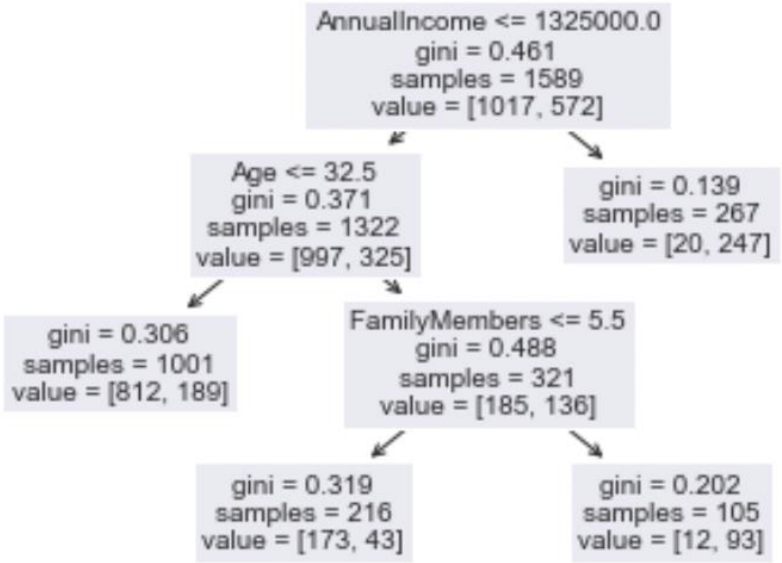
Model Comparison & Observations

- Models with 5 predictors result in the best performance out of the three
- Low sensitivity and high specificity due to imbalance data

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%

Final Model

- Pruned decision trees with three predictors: Annual Income, Age, and Family Members.
- Highest test accuracy out of all models: 82.16%.
- Simplest model to explain to a general audience.



Conclusion

- Check dataset is complete and usable
- Split dataset into training and testing with the ratio 8:2
- Fit different models using training set with different subsets of predictors
- Find the best model based on test accuracy and interpretability

Future Research

- Sampling techniques to address imbalance data
- Other target of interest
- Other potential predictors (Distance, Duration, Reason, Timing, Mode, Region)
- Other possible models (Elastic Net Model, Boosting, Support Vector Machines)

Thank you !