Using Logistic Regression to Predict the Travel Insurance Buyers in India

Jeremy Hampton, Lucy Cobble, Jacob Akubire, Bich Ha Nguyen

Introduction

This paper aims to investigate and identify significant variables that are highly correlated with the choice to purchase travel insurance by customers of the Tour and Travel Company in India. Binary logistic regression is applied on the sample dataset of around 2000 customers in 2019 to find the effective model with the most significant predictors. The model could be used to identify the target customers for travel insurance, thus improving marketing strategies and targeted ads. Cutting down marketing costs could help make the travel insurance price more affordable to other customer segments.

Data Description

The retrieved dataset from Kaggle contains 1,987 observations with 8 predictor variables and 1 response variable.

TravelInsurance: response variable, *the customer's choice to purchase travel insurance from the company*. "1" if the customer bought the travel insurance package during the introductory offering held in the year 2019 and "0" if the customer did not buy it.

Numerical Predictors

- **Age**: years of age of the customer. The range of age is between 25 and 35.
- **AnnualIncome**: yearly income of the customer in rupees, rounded to the nearest 50 thousand.
- **FamilyMembers**: discrete, number of members in the family of the customer.

Categorical Predictors

- **EmploymentType**: customer working sector (government or private/ self-employment)
- **GraduateOrNot**: "Yes" if the customer is a college graduate or "No" if not.
- **FrequentFlyer**: "Yes" means that the customer travels frequently and "No" means that the customer does not. It was obtained based on the customer's history of booking an Air Ticket on at least four different instances within the last 2 years.
- **EverTravelledAbroad**: "Yes" is recorded for customers who have ever travelled abroad and "No" for customers who haven't.
- **ChronicDisease**: This variable indicates if a customer has any chronic diseases or not. "1" if the customer has one or more chronic diseases or "0" if they do not.

Methodology

To predict the probability of a customer purchasing travel insurance, the binary multiple logistic regression is employed. The dataset is split into a training set (80%) and testing set (20%). The analysis begins with the exploratory data analysis (EDA) of all the variables, including their correlations with each other and needs for transformations.The full model is fitted and tested for usefulness and model assumptions. After reviewing the significance of each predictor in the full

model, the variable selection process is applied to produce candidate models with fewer predictors and better fit. PRESS, AIC and BIC are used to select the best main effects model. From the best main effect model, the variable selection process is applied again to produce the best 2-way interaction model. Both models are fitted to the testing dataset to compare their performance through the confusion matrix.

Binary Logistic Regression Model

The response variable assumes Bernoulli distribution which takes values 1 with probability of success (π) or 0 with probability of failure $(1 - \pi)$. Using the logistic regression model, the response variable is transformed to a log scale of odds ratio (the chance of buying travel insurance against not buying travel insurance). The fitted model is:

$$ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = bo + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8$$

$\hat{\pi}$ *- response variable, the probability to purchase travel insurance*

X_1 *- Age;* X_2 *- Employment Type = Private Sector/Self Employed. Government Sector = 0*

X_3 *- GraduateOrNot = Yes. No = 0 ;* X_4 *- Annual Income*

X_5 *- Family Member;* X_6 *- Chronic Disease = Yes. No = 0*

X_7 *- FrequentFlyer = Yes. No = 0;* X_8 *- EverTravelledAbroad = Yes. No = 0*

Model Assumptions

The fitted logistic regression model should satisfy the following assumptions:

1. The response variable follows the Bernoulli distribution, only has two values 0 and 1.
2. Linearity: the log of the odds ratio is linearly related to the predictor variables.
3. Independence: We assume that a customer's decision did not affect another's.
4. No multicollinearity among predictors.
5. Large sample size

For the logistic regression model, non-normal and non-constant variance error could take place as the response variable is binary. Thus the model diagnostics would focus on the model linearity and outliers.

Analysis

EDA

Age, Annual Income and Family Member are fairly right-skewed but there are no outliers so no transformation is needed (Plot 1). In terms of categorical variables, most of customers are graduates (85.2%), working in private sector or self-employed (71.3%), not having any chronic disease (72.2%), not frequent flyer (79%) and never travel abroad before (90.9%) (Plot 2). In terms of response variable, 35% of the respondents actually purchased travel insurance so we have enough data to apply binary logistic regression (Plot 2). Since most variables are categorical, the pairwise plot only indicates the numerical variables are not highly correlated with each other (Plot 3). However, the travel insurance column in the pairwise plot shows some possible interactions with age, annual income, travel abroad and frequent flyer as the plots for not purchasing travel insurance are different from the plot for purchasing travel insurance.

Full Fitted Model

The full multiple logistic regression model is applied to the training dataset and tested for usefulness against a null model (without predictors). The **Likelihood Ratio Goodness-of-Fit test** results in a p-value < 0.05 so we rejected the null hypothesis that none of the predictors is significant (Table 1). **The VIF test for multicollinearity** indicates that all predictors have fairly low values, below 1.5 (Table 2). This suggests that the predictors are not highly correlated and no removal and transformation on predictors is needed. The summary table of the full fitted model shows the **T-tests results for each predictor** (Table 3). The p-values < 0.05 for age, annual income, family member, frequent flyer and ever travel abroad and the corresponding high p-values for employment type, graduate, and chronic disease suggests a variable selection process to produce more useful models with fewer variables. Taking the exponential of the coefficients of the full model and subtracting by 1 shows that age, annual income, family member, frequent flyer and ever travel abroad have a positive effect on the decision to purchase travel insurance (Table 4).

Full Fitted Model Diagnostics

The plot of residuals against estimated probabilities showed a smooth line with zero slope (Plot 4) and the plot of the binned residuals against the predicted values does not have any particular pattern (Plot 5). Thus the full model is a good fit for this data. The half normality plot (Plot 6) also supports this assumption as all the points follow a linear line. The Cook's Distance plot (Plot 7) points out the 3 observations could be influential points. These customers have very good values for the significant predictors but they do not purchase the insurance (Table 5).. We do not investigate whether these observations are due to measurement error or they were correctly recorded. Also, we keep these observations because we don't think these customers are outliers since there could be other factors influencing their decisions but not captured in the dataset.

Variable Selection: Main Effect Model

To come up with candidate models, three approaches are applied to the multiple regression and the null model: forward selection, backward selection, stepwise selection. For each approach, AIC and BIC are used to select the most appropriate model, thus producing 6 models from these methods. However, all methods using AIC points to model (Summary Table 6):

$$TravelInsurance = EverTravelledAbroad + AnnualIncome + FamilyMembers + FrequentFlyer + Age + ChronicDiseases$$

and all methods using BIC points to model (Summary Table 7):

$$TravelInsurance = EverTravelledAbroad + AnnualIncome + FamilyMembers + FrequentFlyer + Age$$

Based on AIC, BIC and PRESS, we prefer the BIC model with 5 predictors because adding an insignificant predictor Chronic Diseases does not make a great impact (Table 8). We conduct model diagnostics for the BIC model and find that model remains a good fit for the data (Plot 8, 9). Plot 10 and 11 shows that there are new influential points based on the main effects BIC model but we decide to keep these observations as they could help explain what the model is not able to predict. The VIF test confirms that there is no multicollinearity within the predictors (Table 9).

Main Effect Model: Coefficient Interpretation (Table 10)

The intercept is meaningless as age could not be 0 in this context. For age, the odds increase by 7.21% for an additional one year increase in age, holding other variables constant. For annual income, the odds increase by 0.00016% for an additional one unit increase in annual income, holding other variables constant. For family members, the odds increase by 20.17% for one additional family member, holding other variables constant. For frequent flyer, the odds are 75.74% higher for customers who often travel than for customers who do not often travel, holding other variables constant. For ever travelled abroad, the odds are 390.1% higher for customers who have ever travelled abroad than for customers who have never travelled abroad, holding other variables constant.

Variable Selection: Interaction Model

Because all predictors in the main effect model are significant, we explore the potentiality of the 2-way interaction model based on the main effect model. We conduct subset regression with exhaustive search and select the model based on AIC (Plot 15). This approach produces the model:

$$TravelInsurance = EverTravelledAbroad{:}\,AnnualIncome + FamilyMembers + FrequentFlyer + Age{:}\,EverTravelledAbroad + FamilyMembers{:}Age + FrequentFlyer : AnnualIncome + FrequentFlyer : FamilyMembers$$

The interaction model has AIC of 1555.7, which is much smaller than its main effect model's AIC of 1670.491. Model diagnostics (Plot 12, 13, 14) indicates the model satisfies the assumptions and fits the data better as the cook's value is much smaller compared to the main effect model. Since the model involves interaction terms, the model coefficients lose the interpretability as the signs for some predictors change compared to the main effect model (Table 11).

Model Performance

The AIC model and the interaction model are selected to be assessed with the test data set. From the confusion matrix for the interaction model, the sensitivity rate is 84.46% and the specificity rate is 81.29% (Plot 16). From the confusion matrix for the BIC model, both the sensitivity and specificity rates are approximately 78% (Plot 17). From these results, it is clear that the interaction model is better in predicting travel insurance buyers at the expense of interpretation.

Results

The analysis showed that age, annual income, family size, whether the customer has ever traveled abroad and how often they travel are good predictors of the customer's choice to buy travel insurance. Based on the residual deviance and AIC values of the two models suggested, the interaction model that includes the interactions between annual income and ever travel abroad, age and ever travel abroad, and age and family members is preferred since it has the lowest residual deviance and AIC value. Also, the sensitivity rate and specificity rate of the interaction model are higher than those of the main effect model, suggesting that it has higher prediction power than the main effect model.

We find that the target customer to buy travel insurance might fit one of the following categories: earns high income, tends to travel frequently and even abroad, around 30s, and has a family. However, our prediction is based on the assumption that the customer population is between 25

and 35 (given in our dataset). To apply this model for future usage, we would recommend a dataset with customers with a wider range of age.

References

Tejashvi. (2021, August 23). *Travel insurance prediction data*. Kaggle. Retrieved 2021, from https://www.kaggle.com/tejashvi14/travel-insurance-prediction-data.

Acknowledgements

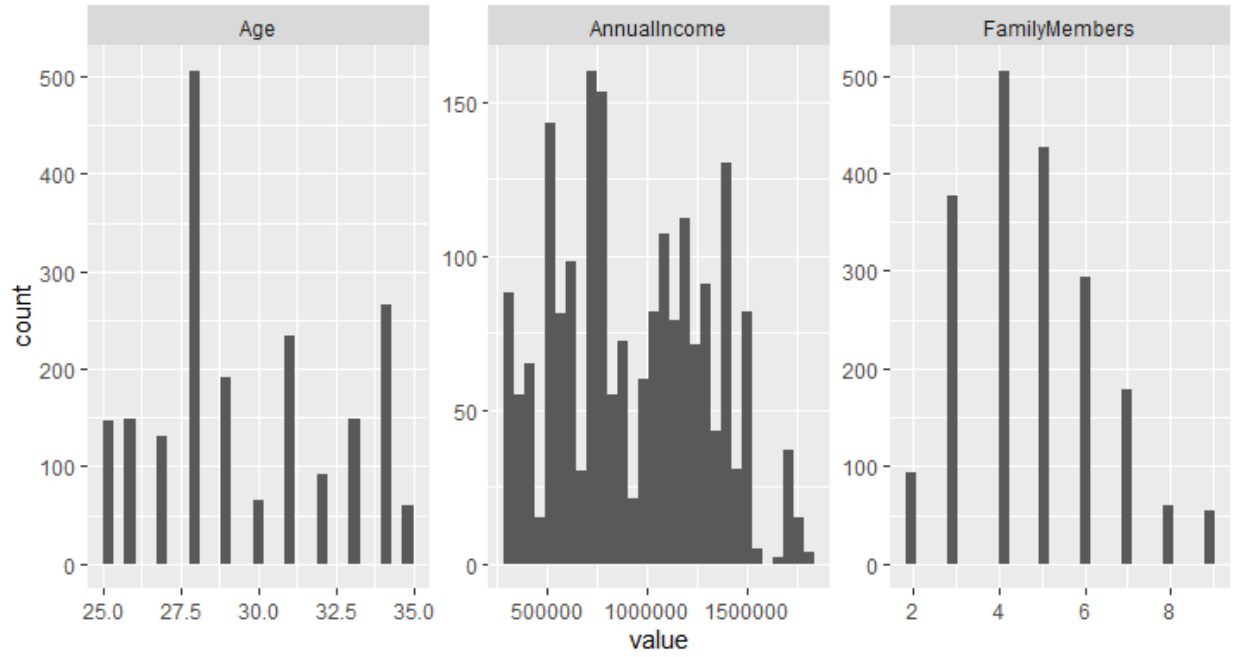
We would like to thank Dr.Yuan and Dr. Milkovich and the TAs for helping us during this report. Also, we would like to thank each individual group member for their individual contributions. Also, without free resources like Kaggle, projects like this would not have been possible.

Another key thing that is important is the open source libraries of R that allowed us to do this project.

Appendix

Plots & Results

Plot 1: Distributions of Age, Annual Income and Family Members



Comment: Age only between 25 and 35, Annual Income also above 250,000. All three variables have fairly right-skewed distributions but still acceptable as there are enough observations for their range of values

Plot 2: Proportion charts for Categorical Variables (taken from the EDA report)

Employment Type character — 0 (0) — 2 (0.001) — - — - — -

Distribution | Character Stats | Cardinality/Missing

Levels	Frequency	Percent
Private Sector/Self Employed	1.417	71,3%
Government Sector	570	28,7%

GraduateOrNot character — 0 (0) — 2 (0.001) — - — - — -

Distribution | Character Stats | Cardinality/Missing

Levels	Frequency	Percent
Yes	1.692	85,2%
No	295	14,8%

ChronicDiseases factor — 0 (0) — 2 (0.001) — - — - — -

Distribution | Cardinality/Missing

Levels	Frequency	Percent
0	1.435	72,2%
1	552	27,8%

FrequentFlyer character — 0 (0) — 2 (0.001) — - — - — -

Distribution | Character Stats | Cardinality/Missing

Levels	Frequency	Percent
No	1.570	79,0%
Yes	417	21,0%

EverTravelledAbroad character — 0 (0) — 2 (0.001) — - — - — -

Distribution | Character Stats | Cardinality/Missing

Levels	Frequency	Percent
No	1.607	80,9%
Yes	380	19,1%

▼	**TravelInsurance** factor	0 (0)	2 (0.001)	-	-	-

Distribution Cardinality/Missing

Levels	Frequency	Percent
0	1.277	64,3%
1	710	35,7%

Plot 3: Pairwise plot of all variables

Table 1: ANOVA - Usefulness of Full Fitted Model Test

```
Analysis of Deviance Table

Model 1: TravelInsurance ~ 1
Model 2: TravelInsurance ~ Age + `Employment Type` + GraduateOrNot + AnnualIncome +
    FamilyMembers + ChronicDiseases + FrequentFlyer + EverTravelledAbroad
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1589     2072.8
2      1581     1655.2  8   417.56 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 2: VIF for Full Fitted Model

Age	`Employment Type`	GraduateOrNot	AnnualIncome	FamilyMembers	ChronicDiseases	FrequentFlyer
1.042063	1.172882	1.062305	1.330567	1.021934	1.007692	1.096083

EverTravelledAbroad
1.152416

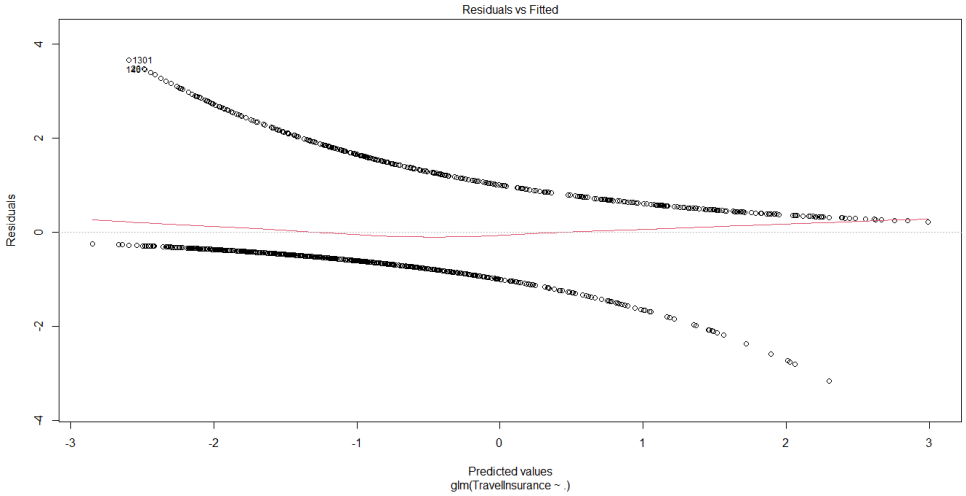
Table 3: Summary of Full Fitted Model

```
Call:
glm(formula = TravelInsurance ~ ., family = binomial(link = logit),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1889  -0.8012  -0.5466   0.7551   2.3087

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -5.555e+00  7.045e-01  -7.885 3.15e-15 ***
Age                                      7.095e-02  2.066e-02   3.434 0.000594 ***
`Employment Type`Private Sector/Self Employed  2.665e-02  1.490e-01   0.179 0.858061
GraduateOrNotYes                        -1.316e-01  1.731e-01  -0.761 0.446926
AnnualIncome                             1.625e-06  1.990e-07   8.166 3.19e-16 ***
FamilyMembers                            1.829e-01  3.769e-02   4.851 1.23e-06 ***
ChronicDiseases1                         2.218e-01  1.343e-01   1.651 0.098714 .
FrequentFlyerYes                         5.671e-01  1.524e-01   3.720 0.000199 ***
EverTravelledAbroadYes                   1.589e+00  1.692e-01   9.393  < 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: 2072.8  on 1589  degrees of freedom
Residual deviance: 1655.2  on 1581  degrees of freedom
AIC: 1673.2

Number of Fisher Scoring iterations: 4
```

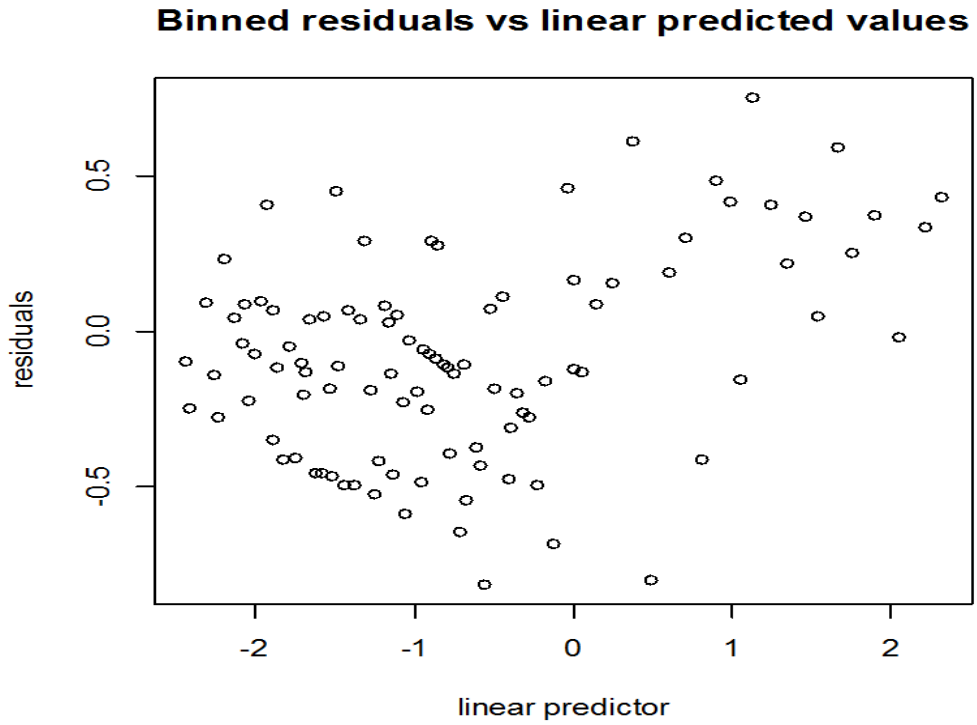
Table 4: (Transformed Coefficients - 1) of Full Fitted Model

(Intercept)	Age	`Employment Type`Private Sector/Self Employed
-9.961322e-01	7.352779e-02	2.701054e-02
GraduateOrNotYes	AnnualIncome	FamilyMembers
-1.233219e-01	1.624825e-06	2.006497e-01
ChronicDiseases1	FrequentFlyerYes	EverTravelledAbroadYes
2.483396e-01	7.631218e-01	3.901083e+00

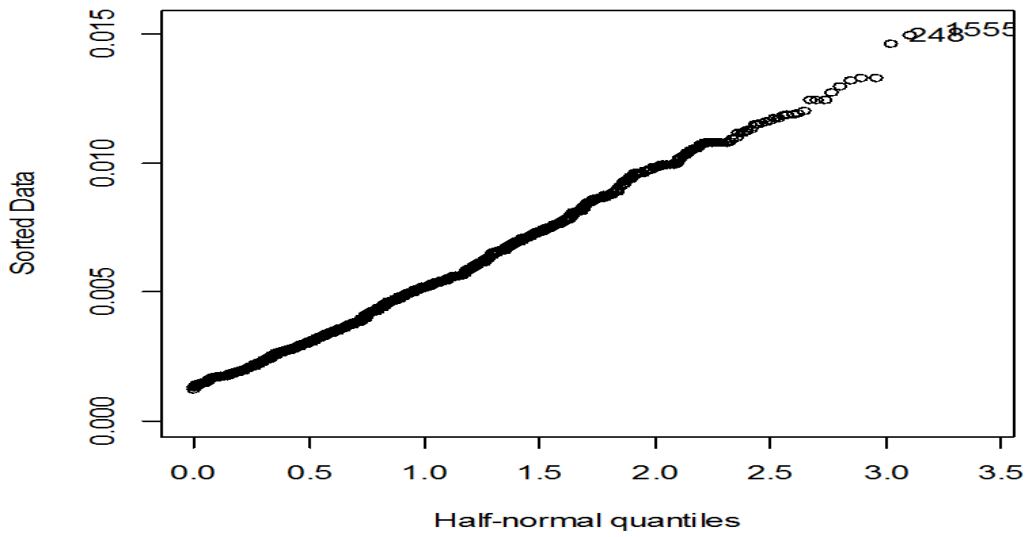
Plot 4: Residuals vs Predicted Probabilities with Lowess Smooth for Full-fitted Model



Plot 5: Binned Residuals vs Linear Predicted Values of Full Fitted Model



Plot 6: Half-normal Probability Plot of Full Fitted Model



Plot 7: Cook's Distance Plot of Full Fitted Model

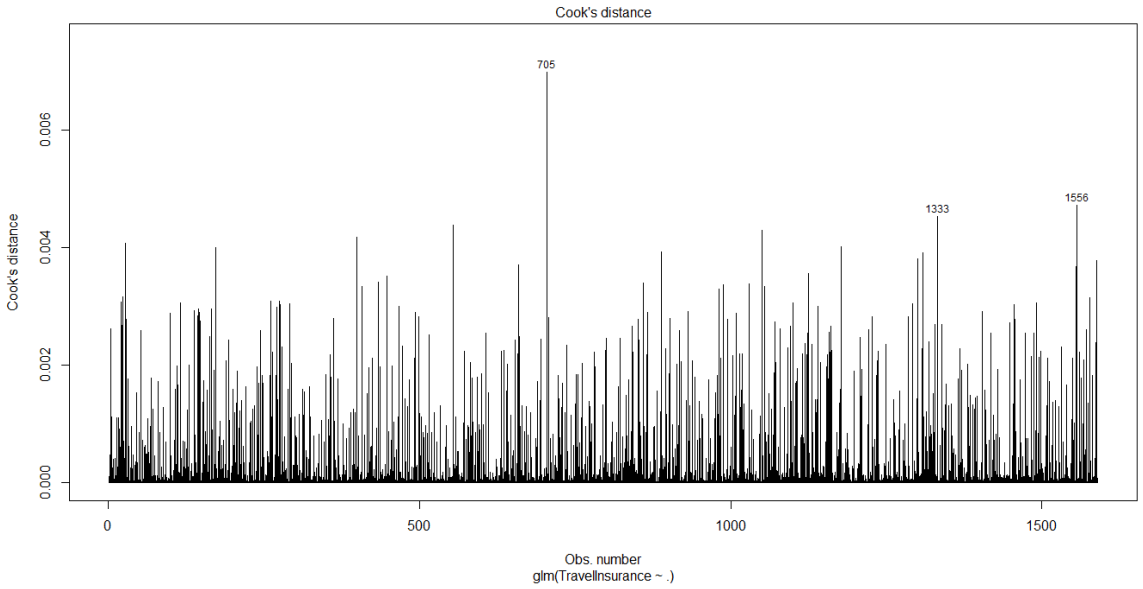


Table 5: Influential Points based on Full-fit model

Age	Employment Type	GraduateOrNot	AnnualIncome	FamilyMembers	ChronicDiseases	FrequentFlyer	EverTravelledAbroad	TravelInsurance
25	Private Sector/Self Employed	No	1700000	5	1	Yes	Yes	0
35	Private Sector/Self Employed	No	950000	6	0	No	Yes	0
34	Private Sector/Self Employed	No	1400000	4	0	No	Yes	0

Table 6: Summary of Main Effects Model based on AIC

```
Call:
glm(formula = TravelInsurance ~ Age + AnnualIncome + FamilyMembers +
    ChronicDiseases + FrequentFlyer + EverTravelledAbroad, family = binomial(link = logit),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1419  -0.8060  -0.5407   0.7590   2.2895

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -5.593e+00  6.814e-01  -8.209 2.23e-16 ***
Age                   6.937e-02  2.045e-02   3.392 0.000695 ***
AnnualIncome          1.622e-06  1.900e-07   8.538  < 2e-16 ***
FamilyMembers         1.820e-01  3.766e-02   4.833 1.35e-06 ***
ChronicDiseases1      2.176e-01  1.342e-01   1.622 0.104813
FrequentFlyerYes      5.773e-01  1.519e-01   3.800 0.000145 ***
EverTravelledAbroadYes  1.585e+00  1.689e-01   9.381  < 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: 2072.8  on 1589  degrees of freedom
Residual deviance: 1655.9  on 1583  degrees of freedom
AIC: 1669.9

Number of Fisher Scoring iterations: 4
```

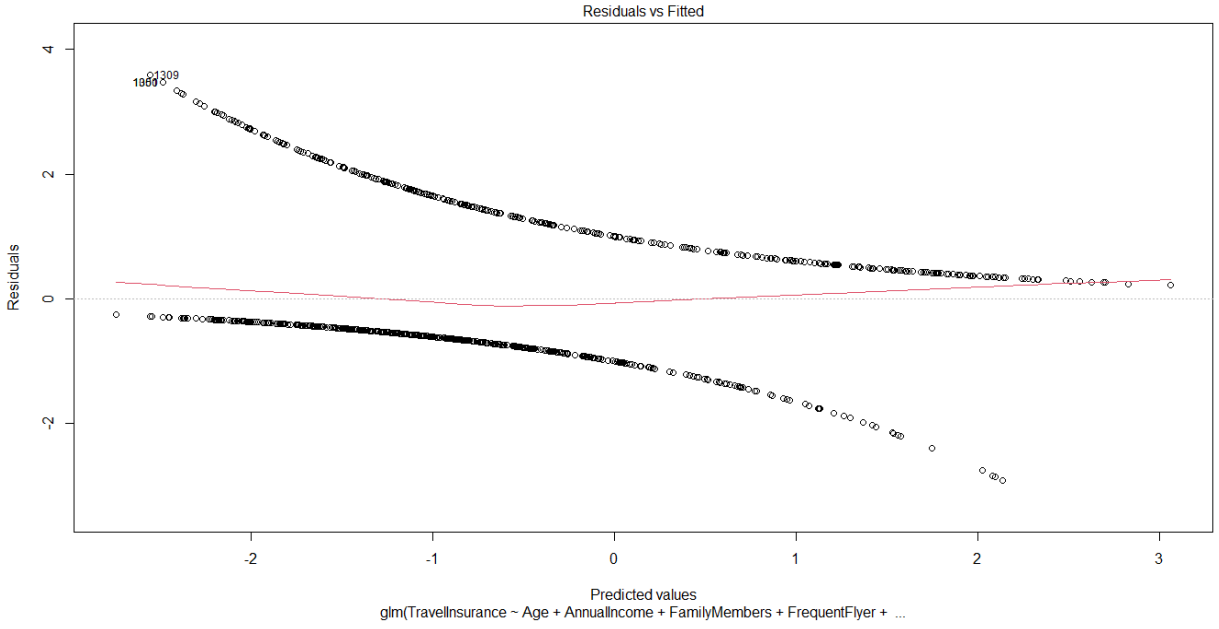
Table 7: Summary of Main Effects Model based on BIC

```
Call:
glm(formula = TravelInsurance ~ Age + AnnualIncome + FamilyMembers +
    FrequentFlyer + EverTravelledAbroad, family = binomial(link = logit),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1201  -0.8028  -0.5474   0.7496   2.2944

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.545e+00  6.792e-01  -8.164 3.24e-16 ***
Age                  6.966e-02  2.042e-02   3.411 0.000647 ***
AnnualIncome         1.620e-06  1.898e-07   8.537  < 2e-16 ***
FamilyMembers        1.837e-01  3.763e-02   4.883 1.04e-06 ***
FrequentFlyerYes     5.638e-01  1.514e-01   3.724 0.000196 ***
EverTravelledAbroadYes  1.590e+00  1.687e-01   9.420  < 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: 2072.8  on 1589  degrees of freedom
Residual deviance: 1658.5  on 1584  degrees of freedom
AIC: 1670.5

Number of Fisher Scoring iterations: 4
```

Table 8: Criterion Values of 2 Main effects models

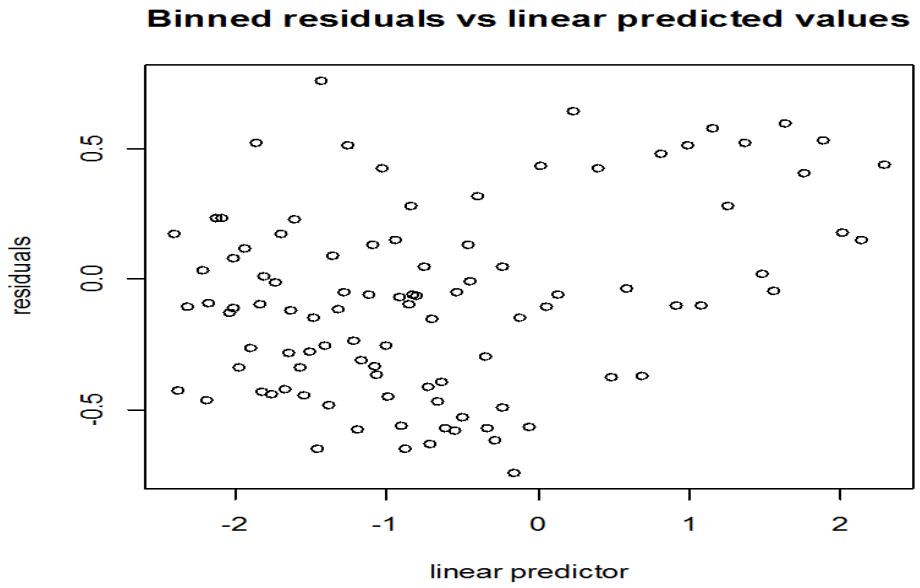
pvec	PRESSvec	AICvec	BICvec
6	1671.417	1669.875	1707.476
5	1671.673	1670.491	1702.720

Plot 8: Residuals vs Predicted Probabilities with Lowess Smooth for Main Effect BIC Model



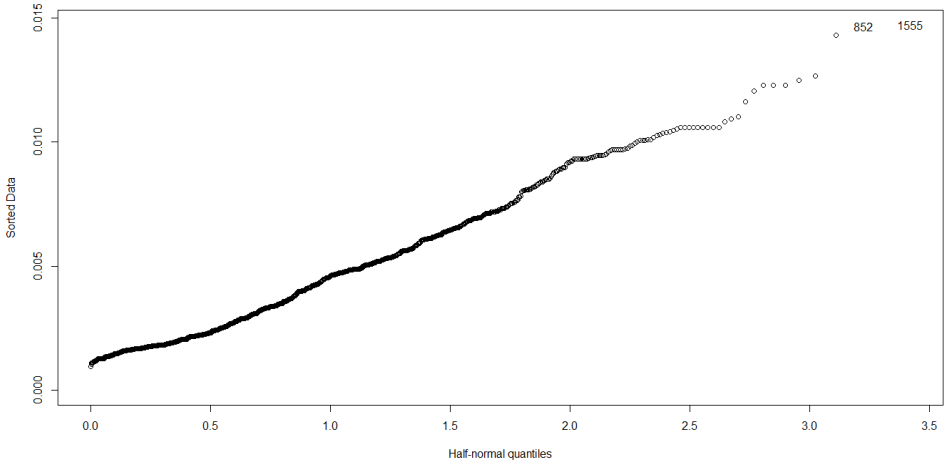
Comment: the smooth line is flat which shows that linearity is met

Plot 9:Binned Residuals vs Linear Predicted Values of Full Fitted Model

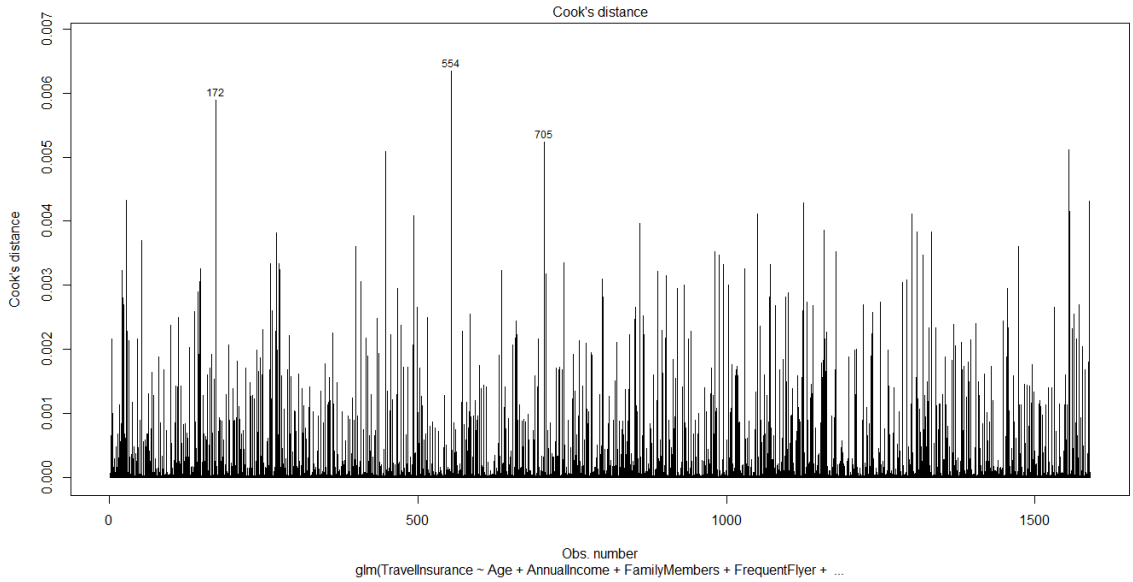


Comment: While the points do not spread out evenly, there is no particular pattern

Plot 10: Half-normal Probability Plot of Main Effects BIC Model



Plot 11:Cook's Distance Plot of Main Effects BIC Model



Comment: 3 possible influential points at 172, 554, 705

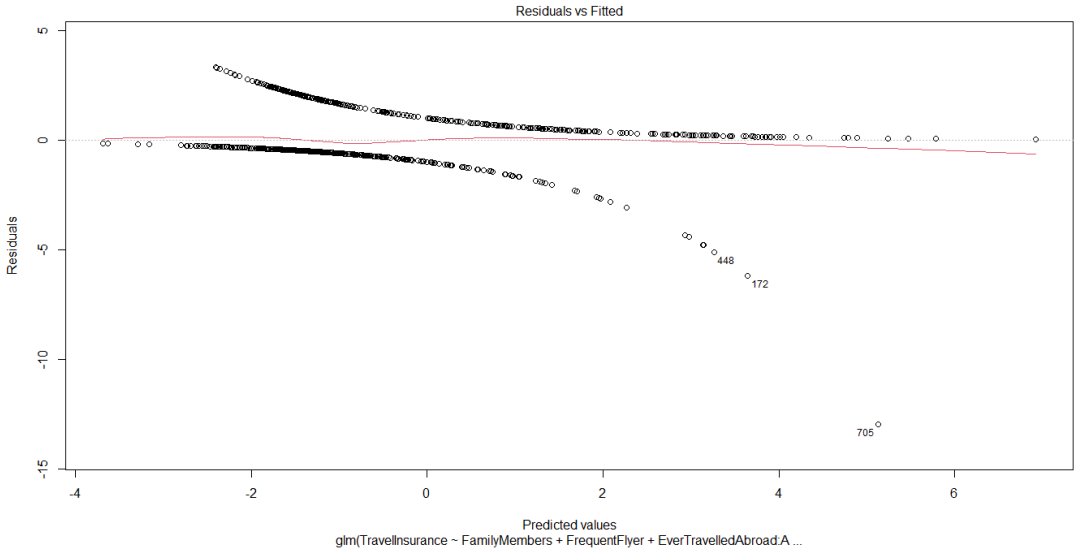
Table 9: VIF of Main Effect BIC model

Age	AnnualIncome	FamilyMembers	FrequentFlyer	EverTravelledAbroad
1.019348	1.213507	1.020157	1.085022	1.148260

Table 10: (Transformed Coefficients - 1) of BIC Main Effect Model

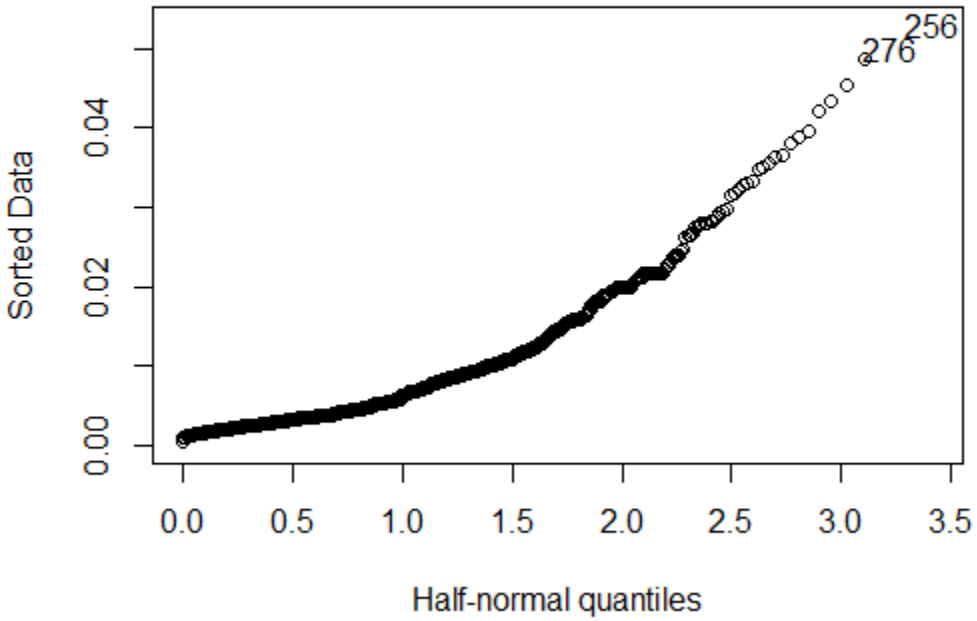
(Intercept)	Age	AnnualIncome	FamilyMembers	FrequentFlyerYes	EverTravelledAbroadYes
-9.960947e-01	7.214643e-02	1.620117e-06	2.017152e-01	7.573973e-01	3.901714e+00

Plot 12: Residuals vs Predicted Probabilities with Lowess Smooth for Interaction Model



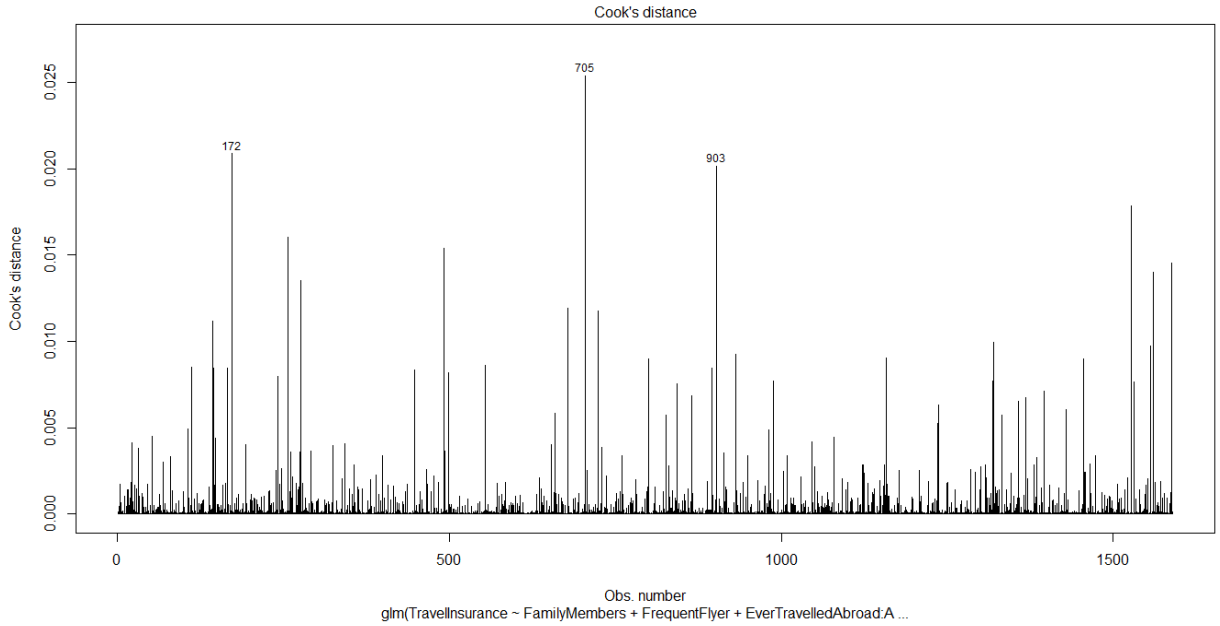
Comment: The smooth line is flat

Plot 13: Half-normal Probability Plot of Interaction Model



Comment: The line seems to be a bit of a curve but still acceptable as all the points follows the line

Plot 14: Cook's Distance Plot of Interaction Model

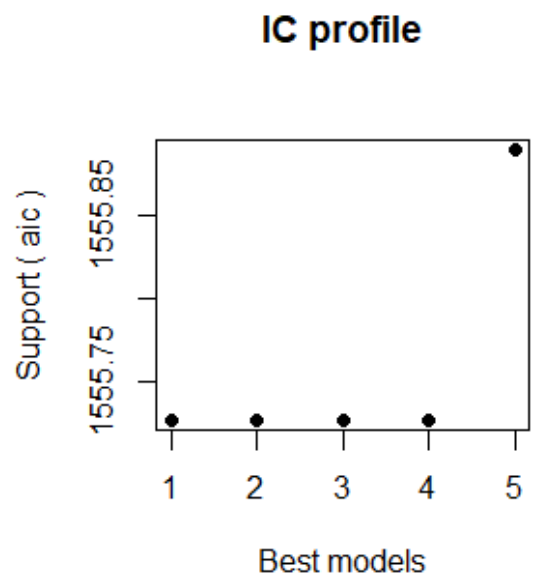


Comment: 3 possible influential points 172, 705, 903

Table 11: Summary Table of Interaction Model

```
Call:
glm(formula = TravelInsurance ~ FamilyMembers + FrequentFlyer +
    EverTravelledAbroad:AnnualIncome + Age:EverTravelledAbroad +
    FamilyMembers:Age + FrequentFlyer:AnnualIncome + FrequentFlyer:FamilyMembers,
    family = binomial(link = logit), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2037  -0.7140  -0.5835   0.6405   2.2324

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        7.851e+00  2.230e+00   3.520 0.000431 ***
FamilyMembers                     -2.702e+00  4.482e-01  -6.030 1.64e-09 ***
FrequentFlyerYes                  -3.154e+00  8.797e-01  -3.586 0.000336 ***
EverTravelledAbroadNo:AnnualIncome  6.224e-07  2.231e-07   2.790 0.005268 **
EverTravelledAbroadYes:AnnualIncome 4.454e-06  6.066e-07   7.343 2.09e-13 ***
EverTravelledAbroadNo:Age         -3.464e-01  7.475e-02  -4.634 3.60e-06 ***
EverTravelledAbroadYes:Age        -4.434e-01  7.906e-02  -5.608 2.04e-08 ***
FamilyMembers:Age                  9.585e-02  1.494e-02   6.416 1.40e-10 ***
FrequentFlyerYes:AnnualIncome      2.719e-06  5.165e-07   5.264 1.41e-07 ***
FamilyMembers:FrequentFlyerYes     1.702e-01  1.170e-01   1.455 0.145663
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2072.8  on 1589  degrees of freedom
Residual deviance: 1535.7  on 1580  degrees of freedom
AIC: 1555.7

Number of Fisher Scoring iterations: 5
```

Plot 15: AIC of 5 Best Subset Models with 2-way interactions



Comment: Out of 5 models, the first 4 have similar AIC, so we choose what model with fewest terms

Table 12: (Transformed Coefficients - 1) of BIC Main Effect Model

(Intercept)	FamilyMembers	FrequentFlyerYes
2.567089e+03	6.705041e-02	4.266948e-02
EverTravelledAbroadNo:AnnualIncome	**EverTravelledAbroadYes:AnnualIncome**	**EverTravelledAbroadNo:Age**
1.000001e+00	1.000004e+00	7.072535e-01
EverTravelledAbroadYes:Age	**FamilyMembers:Age**	**FrequentFlyerYes:AnnualIncome**
6.418440e-01	1.100598e+00	1.000003e+00
FamilyMembers:FrequentFlyerYes		
1.185511e+00		

Plot 16: Confusion Matrix Plot of Interaction Model

	TravelInsurance	
InteractionPredictBuy	**0**	**1**
FALSE	239	55
TRUE	16	87

Comment: The sensitivity rate is calculated as 87/(87 + 16) and specificity rate is calculated as 239/ (239 + 55)

Plot 17: Confusion Matrix Plot of Main Effects Model

	TravelInsurance	
MainPredictBuy	**0**	**1**
FALSE	233	66
TRUE	22	76

Comment: The sensitivity rate is calculated as 76/(76 + 22) and specificity rate is calculated as 233/ (233 + 66)

R Code

```
# Load packages
# install.packages("glmulti")
library(dlookr)
library(caTools)
library(MPV)
library(ggfortify)
library(faraway)
library(car)
library(caret)
library(leaps)
library(glmulti)
library(kableExtra)
```

```
library(GGally)
library(purrr)
library(tidyr)
library(ggplot2)
library(tidyverse)

# Set directory
setwd("C:/Users/77 thaiha/Pictures/2021F/STA 463/Final Project")

# Load data
raw_data = read_csv("TravelInsurancePrediction.csv", col_select=2:10)

# Transform numerical variables to factor variables
raw_data$TravelInsurance <- as.factor(raw_data$TravelInsurance)
raw_data$ChronicDiseases <- as.factor(raw_data$ChronicDiseases)


# Part 1: EDA

# Generate automatic report of all variables
raw_data %>%
  eda_web_report(target = "TravelInsurance", subtitle = "Travel Insurance",
                 output_dir = "./", output_file = "EDA.html", theme = "blue")

# Plot distributions of numerical predictors
raw_data %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram()

# Examine correlation between variables and potential interaction with response variable
ggpairs(raw_data)


#### Comments: ####
# The numerical variables have not too skewed distributions -> no transformation needed
# The categorical variables have fairly good proportions
# The response variable has  ~ 30% of observations as 1 -> enough observations to make
predictions


# Part 2: Split data into training and testing sets - 80/20
```

```
set.seed(101)
sample = sample.split(raw_data$TravelInsurance, SplitRatio = .80)
train = subset(raw_data, sample == TRUE)
test  = subset(raw_data, sample == FALSE)

# Part 3: Full Fitted model
# M0: Null
null_fit <- glm(TravelInsurance ~ 1, data = train, family = binomial (link = logit))
# M1: Full main effects
full_fit <- glm(TravelInsurance ~ ., data=train, family=binomial(link=logit))

# Usefulness of Model Test - M1
anova(null_fit, full_fit, test="LRT")

# VIF
vif(full_fit)

# M1 Summary
summary(full_fit)
# Transformed Coefficients
exp(full_fit$coefficients) - 1

# M1: Diagnostics
# Linearity
plot(full_fit, which=1)
# bin the residuals: 100 bins
train %>%
  mutate(residuals = residuals(full_fit), linpred=predict(full_fit)) %>%
  group_by(cut(linpred, breaks = unique(quantile(linpred, (1:100)/101)))) %>%
  summarise(residuals=mean(residuals), linpred = mean(linpred)) %>%
  ggplot() +
  geom_point(aes(x = linpred, y=residuals)) +
  labs(x="linear predictor", title = "Binned Residuals vs linear predicted values")
# halfnormal plot
halfnorm(hatvalues(full_fit))
# influential points
plot(full_fit, which=4)
train[c(705, 1333, 1556),] %>% kbl() %>% kable_styling()


## Part 4: Variable Selection for Main effect model

# AIC Backward
back_fit_AIC <- stats::step(full_fit, direction = "backward", trace=0, plot=T)
```

```
summary(back_fit_AIC)

# Age + Annual Income + Family + Chronic + Frequent + EverTravel

# BIC Backward
back_fit_BIC <- stats::step(full_fit, direction = "backward", k= log(nrow(train)), trace=0)
summary(back_fit_BIC)
# Age + Annual Income + Family + Frequent Flyer + Travel Abroad

# AIC Forward
forw_fit_AIC=step(null_fit, direction="forward", scope=list(upper=full_fit), trace = 0)
summary(forw_fit_AIC)
# same as backward

# BIC Forward
forw_fit_BIC=step(null_fit, direction="forward", k= log(nrow(train)),
scope=list(upper=full_fit), trace = 0)
summary(forw_fit_BIC)
# same as backward

# Stepwise Selection
stepwise_AIC=step(null_fit, direction="both", scope=list(upper=full_fit),trace=0)
summary(stepwise_AIC)
# same as backward

stepwise_BIC=step(null_fit, direction="both",
scope=list(upper=full_fit),k=log(nrow(train)), trace=0)
summary(stepwise_BIC)
# same as backward

## Compare the best main effects model

# M2: Age + Annual Income + Family + Chronic + Frequent + EverTravel

# M3: Age + Annual Income + Family + Frequent Flyer + Travel Abroad
AICvec=c(AIC(back_fit_AIC),AIC(back_fit_BIC)) # AIC citerion
BICvec=c(BIC(back_fit_AIC),BIC(back_fit_BIC)) # BIC criterion
PRESSvec=c(PRESS(back_fit_AIC),PRESS(back_fit_BIC)) # PRESS Criterion
pvec=c((summary(back_fit_AIC)$df[1]-1),(summary(back_fit_BIC)$df[1]-1))

data=cbind(pvec,PRESSvec,AICvec,BICvec)
data %>% kbl() %>% kable_styling()

# Comment: Not really a huge improvement with model including the Chronic Disease
```

```
# Furthermore, the T-test result shows that Chronic Disease is not meaningful so we
decide to exclude


## Best main effect model - BIC_back_fit model
# Summary table
summary(back_fit_BIC)
# Diagnostics
plot(back_fit_BIC, which = 1)
# bin the residuals: 100 bins
train %>%
  mutate(residuals = residuals(back_fit_BIC), linpred=predict(back_fit_BIC)) %>%
  group_by(cut(linpred, breaks = unique(quantile(linpred, (1:100)/101)))) %>%
  summarise(residuals=mean(residuals), linpred = mean(linpred)) %>%
  ggplot() +
  geom_point(aes(x = linpred, y=residuals)) +
  labs(x="linear predictor", title = "Binned Residuals vs linear predicted values")
# half normal
halfnorm(hatvalues(back_fit_BIC))
# influential
plot(back_fit_BIC, which=4)
# VIF
vif(back_fit_BIC)

# Coefficient Interpretation
exp(back_fit_BIC$coefficients) - 1 # percent change


## Part 5: Fit the main interaction model
glmulti.logistic.out <- do.call("glmulti",
       list(TravelInsurance ~ AnnualIncome + EverTravelledAbroad + Age + FamilyMembers
+ FrequentFlyer,
         data = train,
         level = 2,              # 2 way interaction considered
         method = "h",           # Exhaustive approach
         crit = "aic",           # AIC as criteria
         confsetsize = 5,        # Keep 5 best models
         plotty = T, report = T,
         fitfunction = "glm",
         family = binomial))     # lm function

## Show 5 best models (Use @ instead of $ for an S4 object)
glmulti.logistic.out@formulas
```

```
## Model with the least number of predictors
summary(glmulti.logistic.out@objects[[2]])

## Results from console ######################
# FamilyMembers                    -2.702e+00
# FrequentFlyerYes                 -3.154e+00
# EverTravelledAbroadNo:AnnualIncome   6.224e-07
# EverTravelledAbroadYes:AnnualIncome  4.454e-06
# EverTravelledAbroadNo:Age          -3.464e-01
# EverTravelledAbroadYes:Age         -4.434e-01
# FamilyMembers:Age                  9.585e-02
# FrequentFlyerYes:AnnualIncome       2.719e-06
# FamilyMembers:FrequentFlyerYes
#############################################

# Using the summary, fit the interaction model
# M4: best interaction model based on AIC
best_inter <- glm(TravelInsurance ~ FamilyMembers + FrequentFlyer +
EverTravelledAbroad:AnnualIncome +
                  Age:EverTravelledAbroad + FamilyMembers:Age +
FrequentFlyer:AnnualIncome +
                  FrequentFlyer:FamilyMembers, data = train, family = binomial (link = logit))
# Diagnostics
# Linearity
plot(best_inter, which =1)
halfnorm(hatvalues(best_inter))
# Influential
plot(best_inter, which = 4)
# Summary
summary(best_inter)
# Coefficient Interpretation
exp(best_inter$coefficients) # odds
# Comment: The coefficients change compared to the main effect model

## Part 5: Model Performance on Testing data

# Interaction model
test_trial <- test %>%
  mutate(inter_prob = predict(best_inter, newdata=test,
                              type="response"),
         InteractionPredictBuy = inter_prob >= 0.5,
         BIC_prob = predict(back_fit_BIC, newdata=test,
                            type="response"),
         MainPredictBuy = BIC_prob >= 0.5)
```

```
# Confusion Matrix
# NOTE: True = Predict Buy, False = Predict Not Buy, TravelInsurance: 0 = Actually Not Buy, 1 = Actually Buy

# Interaction Model:
xtabs(~InteractionPredictBuy + TravelInsurance, data=test_trial)
# Main Effect Model
xtabs(~MainPredictBuy + TravelInsurance, data=test_trial)
```