Latent Analytics Pvt. Ltd.

Presented by
Sanjeev Ramasamy Seenivasagamani, Woojin Park, Xiaoyan Zhang, Jeet Ganatra, Christy Sato and Aniruddh Garge

Introduction

- As an analytics solutions provider, we help airline companies make informed decisions with the help of our granular insights
- Analyzing the survey responses and establishing the determining factors for the best customer satisfaction
- Identifying the important business questions and answering them appropriately
- Finding out the best predictors to increase customer satisfaction using various analytical models

Business Questions

- How can airline companies improve customer satisfaction?

- What are the characteristics (class, gender, and age) of customers who are most likely to rate satisfaction higher?

Solutions

- Find good predictors for customer satisfaction (Linear Modeling).
- Generate plots for significant variables and then perform analysis on how to have a higher satisfaction rate.
- Segregate data that is Satisfied and not Satisfied (Support Vector Machines).
- Sorting the evidence and making sense of it (Association Rules Algorithm).

Descriptive Statistics

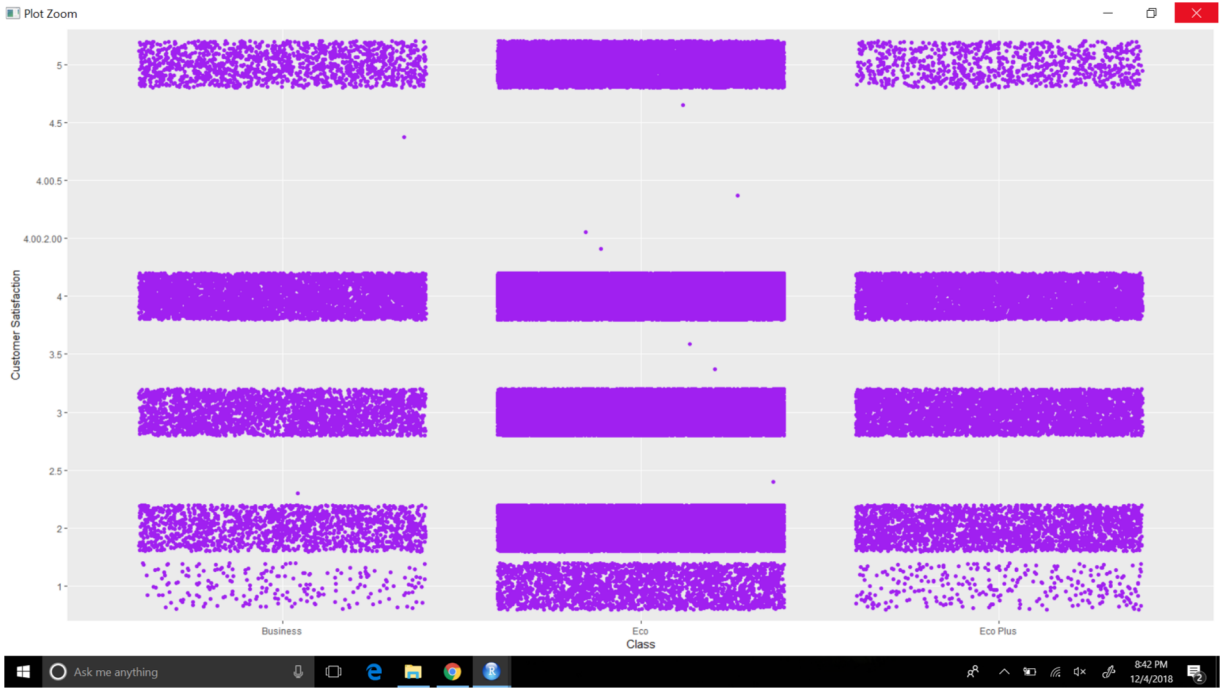
Variable	Data Type	Description
Satisfaction	Numerical	Rated between 1-5 (5 means higher satisfied, and 1 is lowest level of satisfaction)
Airline Status	Categorical	Each customer has a different type of airline status which are platinum, gold, silver, or blue
Age	Numerical	Specific customer's age, starting from 15 to 85 years old
Gender	Categorical	Female/Male
Price Sensitivity	Numerical	Grade to which the price affects customer's purchasing (range between 0-5)
Year of First Flight	Numerical	First flight for each single customer ranging from 2003 to 2012
Number of Flights	Numerical	Number of flights that each customer has taken, range from 0-100
Type of Travel	Categorical	Three traveling purposes (business, mileage tickets based on loyalty cards, or personal travel)

Descriptive Statistics

Variable	Data Type	Description
Flight Cancelled	Categorical	If airline does not operate the flight
Flight time in minutes	Numerical	Period time to destination
Flight Distance	Numerical	Distance between destinations, ranging from 31 to 4983 minutes
Arrival Delay greater than 5 minutes	Categorical	Delay of arrival airline time, in which more than 5 minutes per each customer
Shopping Amount at Airport	Numerical	Three different kinds of service (business, economy plus, or economy)
Class	Categorical	Three different kinds of service (business, economy plus, or economy)

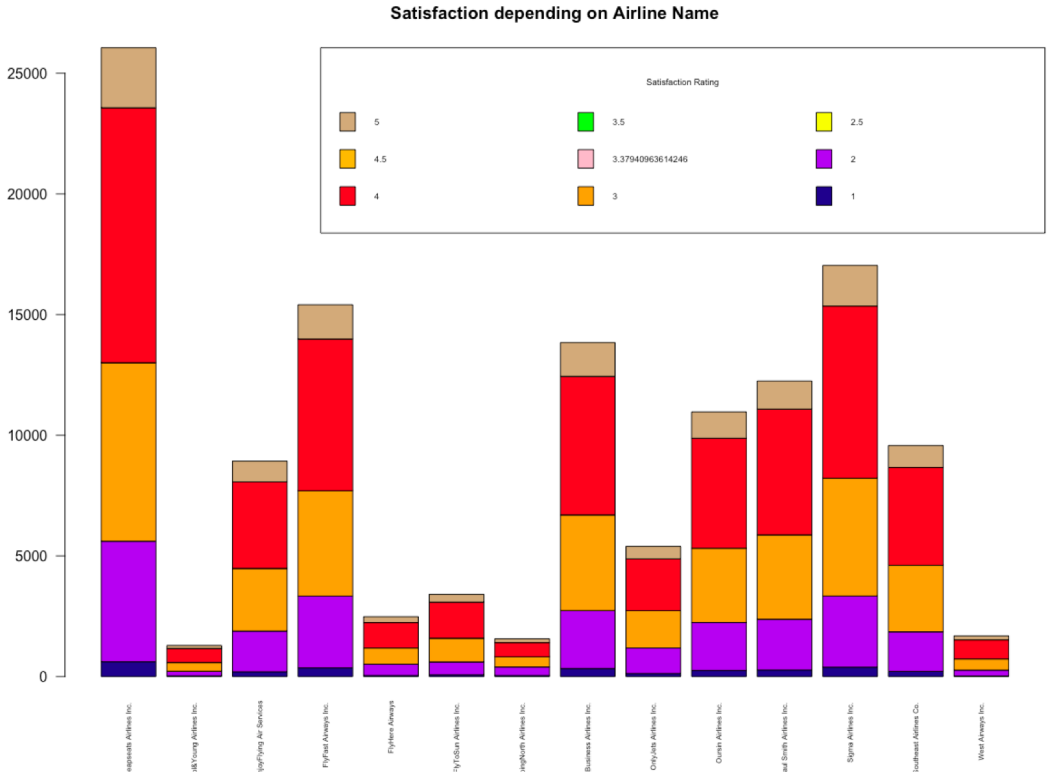
Variable Plots:

1. Class vs Customer Satisfaction



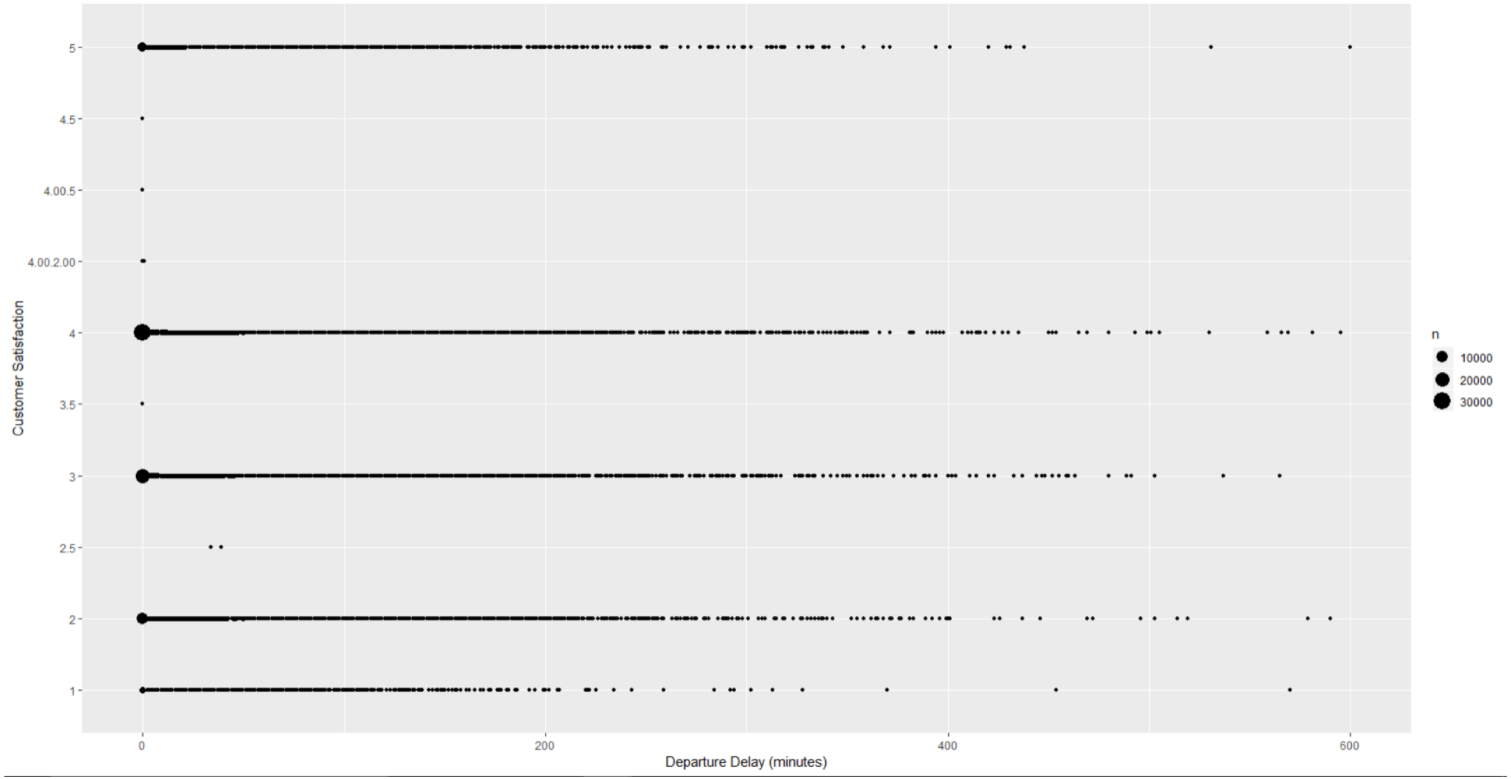
Variable Plots:

2. Airline Name vs Customer Satisfaction



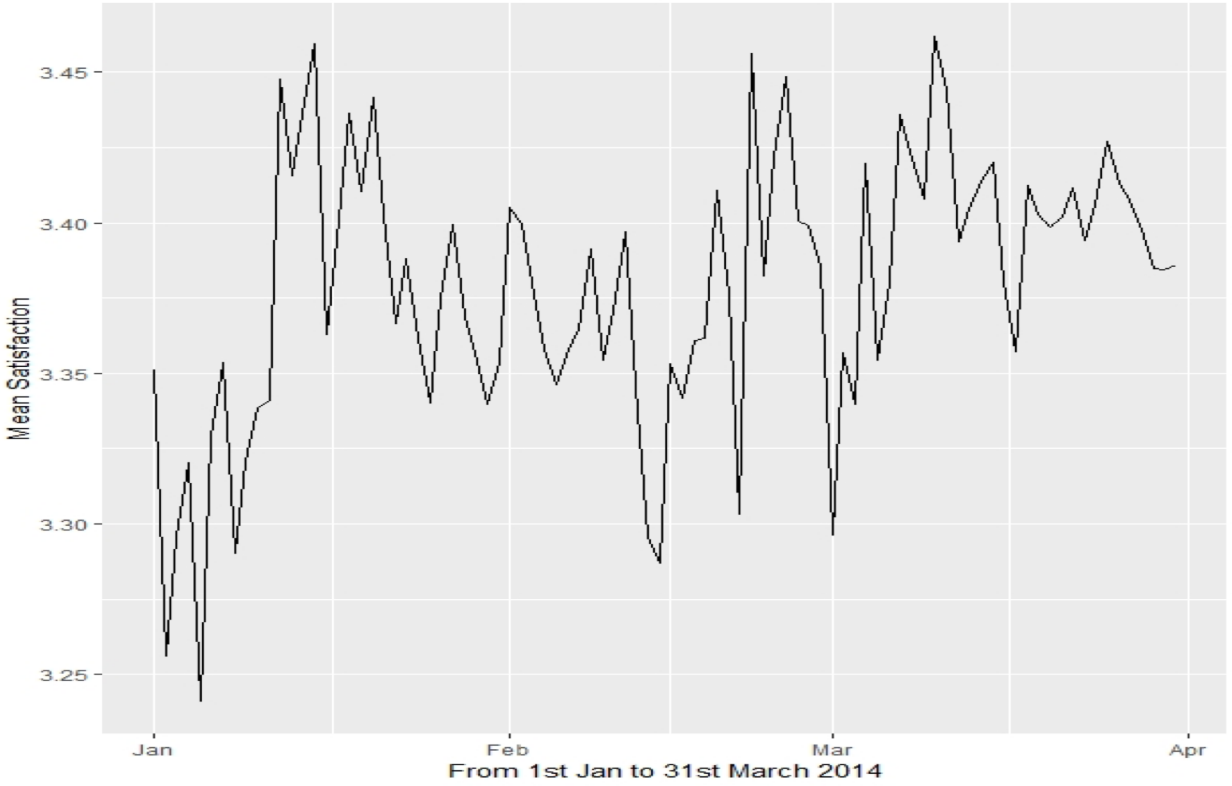
Variable Plots:

3. Departure Delay vs Customer Satisfaction



Variable Plots:

3. Flight Date vs Customer Satisfaction



Linear Model

```
Call:
lm(formula = Satisfaction ~ Airline.Status + Type.of.Travel +
    Arrival.Delay.greater.5.Mins, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8284 -1.0550 -0.3522  0.9449  7.5721

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.543176   0.008203  675.79   <2e-16 ***
Airline.StatusGold              1.297103   0.017980   72.14   <2e-16 ***
Airline.StatusPlatinum          1.285198   0.027968   45.95   <2e-16 ***
Airline.StatusSilver            1.511788   0.012513  120.81   <2e-16 ***
Type.of.TravelMileage tickets  -0.450490   0.018680  -24.12   <2e-16 ***
Type.of.TravelPersonal Travel  -2.488116   0.010885 -228.57   <2e-16 ***
Arrival.Delay.greater.5.Minsyes -0.627200   0.010292  -60.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.76 on 129882 degrees of freedom
Multiple R-squared:  0.3947,    Adjusted R-squared:  0.3946
F-statistic: 1.411e+04 on 6 and 129882 DF,  p-value: < 2.2e-16
```

Linear Model

```
Call:
lm(formula = Satisfaction ~ Airline.Status + Type.of.Travel +
    Arrival.Delay.greater.5.Mins, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8284 -1.0550 -0.3522  0.9449  7.5721

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.543176   0.008203  675.79   <2e-16 ***
Airline.StatusGold              1.297103   0.017980   72.14   <2e-16 ***
Airline.StatusPlatinum          1.285198   0.027968   45.95   <2e-16 ***
Airline.StatusSilver            1.511788   0.012513  120.81   <2e-16 ***
Type.of.TravelMileage tickets  -0.450490   0.018680  -24.12   <2e-16 ***
Type.of.TravelPersonal Travel  -2.488116   0.010885 -228.57   <2e-16 ***
Arrival.Delay.greater.5.Minsyes -0.627200   0.010292  -60.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.76 on 129882 degrees of freedom
Multiple R-squared:  0.3947,     Adjusted R-squared:  0.3946
F-statistic: 1.411e+04 on 6 and 129882 DF,  p-value: < 2.2e-16
```

Support Vector Machine

Check Point

- Percentage of confusion matrix group :Not Satisfied(49%), Satisfied(51%)
- This model enable to predict almost with 80% accuracy
- Interesting finding : Overall error rate is 0.24%
- But, when we compare the error rate of two groups
- Satisfied group Error rate : 2563/(19568+2563) = 0.11%
- Not Satisfied group Error rate : 7693/(7693+13473) = 0.36 %
- Better accuracy(89%) to distinguish ‘Satisfied’ customers (*Not Satisfied 64%)

```
> str(newtrainData)
'data.frame':   86592 obs. of  8 variables:
 $ Age                          : int  71 46 48 42 69 31 63 57 56 41 ...
 $ Gender                       : Factor w/ 2 levels "Female","Male": 1 1 2 2 1 1 2 2 1 1 ...
 $ Price.Sensitivity            : int  1 1 1 2 0 2 1 2 0 1 ...
 $ No.of.Flights.p.a.           : int  64 2 0 32 18 12 21 21 18 1 ...
 $ Type.of.Travel               : Factor w/ 3 levels "Business travel",..: 3 1 1 3 1 3 1 1 1 2 ...
 $ Class                        : Factor w/ 3 levels "Business","Eco",..: 2 2 2 2 3 2 3 2 2 2 ...
 $ Arrival.Delay.greater.5.Mins: Factor w/ 2 levels "no","yes": 2 2 2 1 2 1 2 2 2 1 ...
 $ overallSatisfaction          : chr  "Not Satisfied" "Satisfied" "Satisfied" "Satisfied" ...
```

```
> ksvm(overallSatisfaction ~., data=newtrainData, kernel = "rbfdot",kpar="automatic", C=5,cross=3, prob.m
odel=TRUE)
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.206272679803563

Number of Support Vectors : 40013

Objective Function Value : -191525.6
Training error : 0.214593
Cross validation error : 0.217249
Probability model included.
```

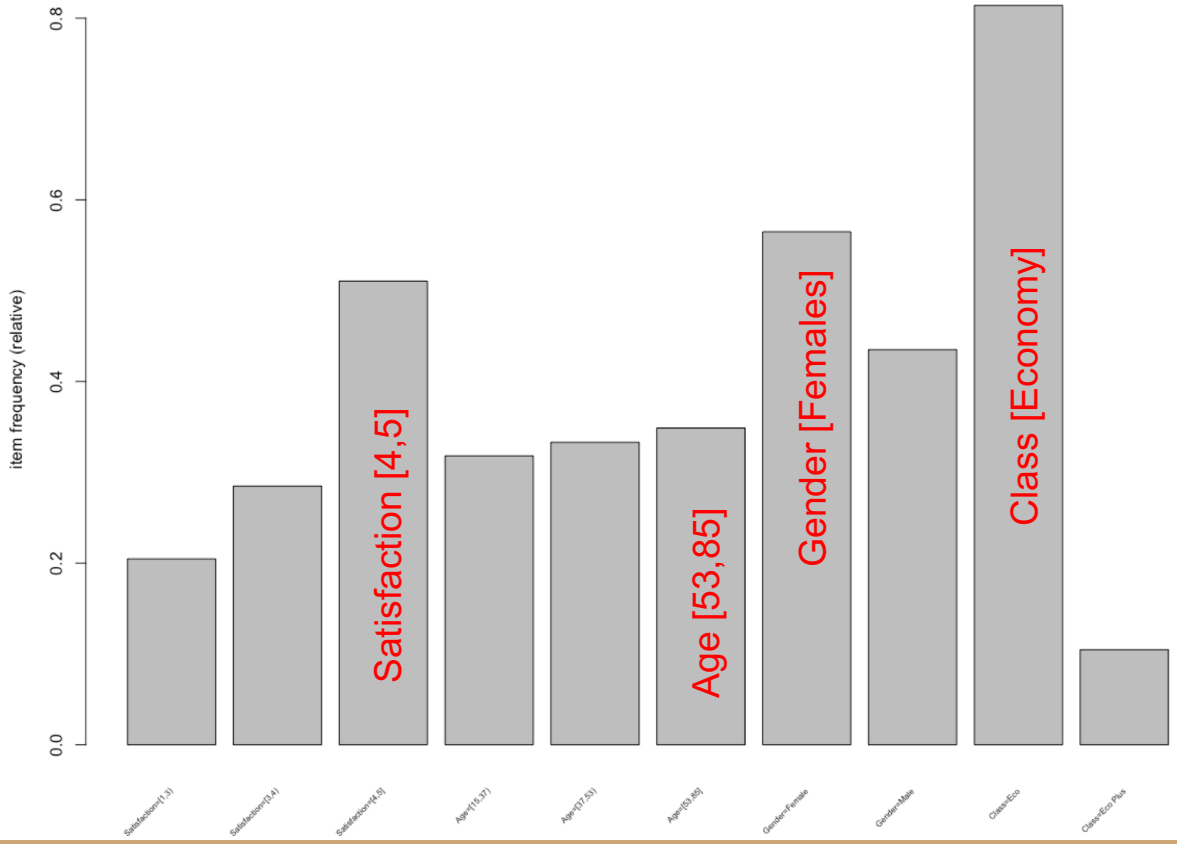
Support Vector Machine

Lesson-learned

- SVM is very good when we have no idea on the data
- Also it has a regularisation parameter (finding optimal margin) to avoid over-fitting or under-fitting problem
- Need to be patient while building SVMs on large datasets, they take a lot of time for training

```
> table(comTable)
                svmPred.1...
newtestData...8.     0     1
   Not Satisfied  7693 13473
   Satisfied     19568  2563
> # 11. Calculate an error rate based on what you see in the confusion matrix.
> t<-table(comTable)
> sum(t[1,1]+t[2,2])/sum(t)
[1] 0.2368755
```

Association Rules Mining



Association Rules Mining

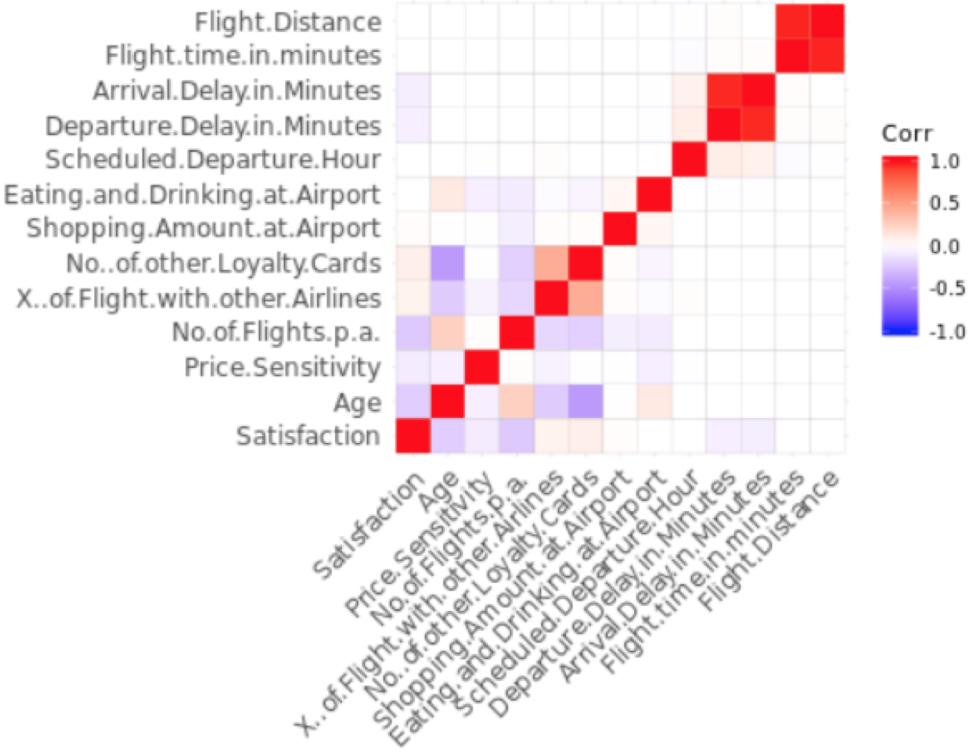
High Satisfaction Ratings

	lhs		rhs	support	confidence	lift	count
[1]	{Age=[37,53),Gender=Male,Class=Business}	=>	{Satisfaction=[4,5]}	0.01134815	0.7967568	1.560652	1474
[2]	{Age=[37,53),Class=Business}	=>	{Satisfaction=[4,5]}	0.02138749	0.7335622	1.436869	2778
[3]	{Age=[37,53),Gender=Male}	=>	{Satisfaction=[4,5]}	0.11217270	0.7321976	1.434196	14570
[4]	{Age=[37,53),Gender=Male,Class=Eco}	=>	{Satisfaction=[4,5]}	0.09536604	0.7257441	1.421555	12387
[5]	{Age=[37,53),Gender=Female,Class=Business}	=>	{Satisfaction=[4,5]}	0.01003934	0.6732060	1.318646	1304
[6]	{Age=[37,53),Class=Eco Plus}	=>	{Satisfaction=[4,5]}	0.01964755	0.6647564	1.302095	2552

Low Satisfaction Ratings

	lhs		rhs	support	confidence	lift	count
[1]	{Age=[53,85],Gender=Female,Class=Eco}	=>	{Satisfaction=[1,3)}	0.05668686	0.3570805	1.744427	7363
[2]	{Age=[53,85],Gender=Female}	=>	{Satisfaction=[1,3)}	0.07233099	0.3522817	1.720984	9395
[3]	{Age=[53,85],Class=Eco}	=>	{Satisfaction=[1,3)}	0.09536604	0.3393234	1.657679	12387
[4]	{Age=[53,85],Gender=Female,Class=Eco Plus}	=>	{Satisfaction=[1,3)}	0.01100170	0.3384652	1.653487	1429
[5]	{Age=[53,85]}	=>	{Satisfaction=[1,3)}	0.11703839	0.3356072	1.639525	15202
[6]	{Age=[53,85],Class=Eco Plus}	=>	{Satisfaction=[1,3)}	0.01381949	0.3311197	1.617602	1795

Bivariate Correlation



Logistic Regression

- The dependent variable is converted into a categorical variable based on its value
- If the Satisfaction rate is greater than 3, a value of 1 is assigned to the new ‘Sat’ categorical else 0
- The continuous version of the Satisfaction variable is removed from the dataset
- To check the variable significance, we run the logistic regression model on the dataset and remove few insignificant variables
- The data is split into train & test dataset and the former is used to train the model
- K-fold cross validation is used to validate the model with k=10

Logistic Regression (cont.)

```
Coefficients:
                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                    1.5406746  0.0576884   26.707  < 2e-16 ***
Airline.StatusGold             0.9128970  0.0325792   28.021  < 2e-16 ***
Airline.StatusPlatinum         0.5060946  0.0475544   10.642  < 2e-16 ***
Airline.StatusSilver           1.7253104  0.0256039   67.385  < 2e-16 ***
Age                           -0.0074138  0.0006225  -11.909  < 2e-16 ***
GenderMale                     0.3392020  0.0178278   19.027  < 2e-16 ***
Price.Sensitivity             -0.1650692  0.0161152  -10.243  < 2e-16 ***
No.of.Flights.p.a.            -0.0132780  0.0006667  -19.915  < 2e-16 ***
`Type.of.TravelMileage tickets` -0.4036537  0.0285939  -14.117  < 2e-16 ***
`Type.of.TravelPersonal Travel` -3.0459892  0.0254125 -119.862  < 2e-16 ***
No..of.other.Loyalty.Cards    -0.0563079  0.0083329   -6.757 1.41e-11 ***
ClassEco                      -0.3006609  0.0323818   -9.285  < 2e-16 ***
`ClassEco Plus`               -0.3493020  0.0411090   -8.497  < 2e-16 ***
Scheduled.Departure.Hour       0.0197443  0.0018953   10.418  < 2e-16 ***
Flight.cancelledYes           -0.7543256  0.0676416  -11.152  < 2e-16 ***
Arrival.Delay.greater.5.Minsyes -0.9675506  0.0185289  -52.219  < 2e-16 ***
---
```

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 14883  3349
         1  6353 18712

               Accuracy : 0.7759
                 95% CI : (0.772, 0.7798)
    No Information Rate : 0.5095
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5505
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7008
            Specificity : 0.8482
         Pos Pred Value : 0.8163
         Neg Pred Value : 0.7465
             Prevalence : 0.4905
         Detection Rate : 0.3437
   Detection Prevalence : 0.4211
      Balanced Accuracy : 0.7745
```

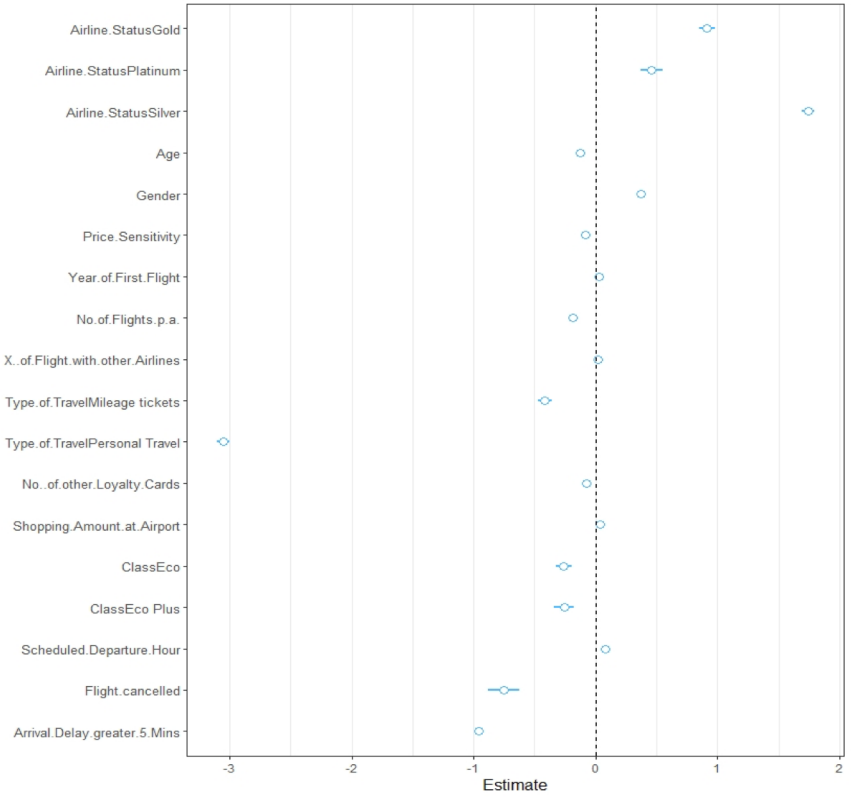
Logistic Regression (cont.)

	Overall
`Type.of.TravelPersonal Travel`	100.000
Airline.StatusSilver	53.603
Arrival.Delay.greater.5.Minsyes	40.194
Airline.StatusGold	18.800
No.of.Flights.p.a.	11.633
GenderMale	10.848
`Type.of.TravelMileage tickets`	6.507
Age	4.555
Flight.cancelledYes	3.885
Airline.StatusPlatinum	3.435
Scheduled.Departure.Hour	3.236
Price.Sensitivity	3.082
ClassEco	2.235
`ClassEco Plus`	1.538
No..of.other.Loyalty.Cards	0.000

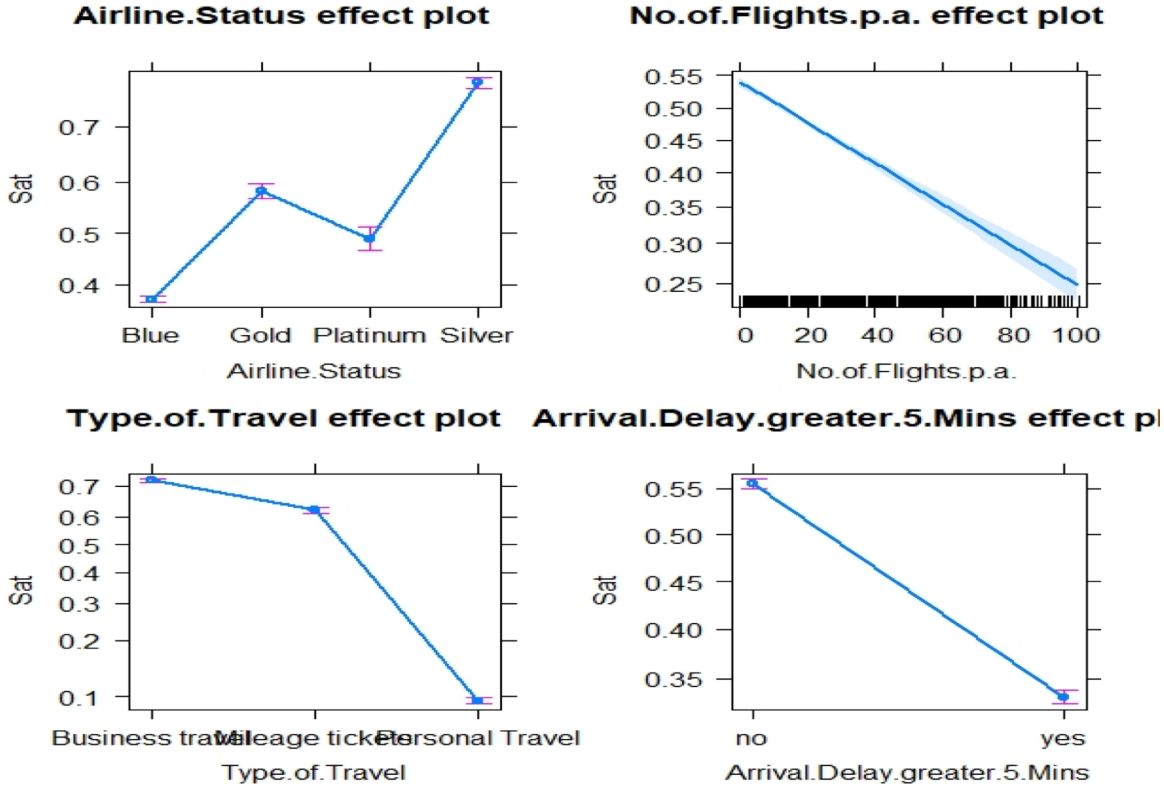
	Overall
`Type.of.TravelPersonal Travel`	100.000
Airline.StatusSilver	47.742
Arrival.Delay.greater.5.Minsyes	31.944
Airline.StatusGold	13.474
No.of.Flights.p.a.	5.886
`Type.of.TravelMileage tickets`	3.319
Airline.StatusPlatinum	0.000

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 14357  3293
         1  6662 18985

               Accuracy : 0.7701
                 95% CI : (0.7661, 0.774)
    No Information Rate : 0.5145
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5377
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6830
            Specificity : 0.8522
         Pos Pred Value : 0.8134
         Neg Pred Value : 0.7402
             Prevalence : 0.4855
         Detection Rate : 0.3316
   Detection Prevalence : 0.4076
      Balanced Accuracy : 0.7676
```

Logistic Regression (cont.)



Logistic Regression (cont.)



Final Conclusions

- Create a strategy to improve overall customer satisfaction within women between the ages of 53 to 85 that are flying in the economy class.
- The class in which the customers travel are significant but not the deciding factor.
- Gender plays a significant role in determining the customer’s satisfaction.
- Lower the delay, higher the customer satisfaction.
- Airline Status - Silver has been predicted to have the highest satisfaction rate followed by Gold.

Thank you.