AIRLINE PASSENGER SATISFACTION

Data Mining Course Project

Group 10

- Safoora Naureen (1002050531)
- Dumpa Akash Reddy (1002025351)
- Spandana Sanapureddy (1002102072)
- Rakesh Maram (1002024578)

Business Context & Problem Statement

Business Context

- As the COVID cases seem to decrease in the past few months, many people are turning towards travel more than in pre-COVID times. The airline company wants to take advantage of the situation and attain the maximum number of passengers.

Business Problem Statement

- •The airline company, wants the data analysis team to analyze and predict which facilities of airlines can the company make changes to, so that there is an increase in business (4-5 times) class passengers, without effecting the economy class.

Feedback is always the best way to measure customer satisfaction and analyze the various factors where we can improve the business. Predicting customer satisfaction through feedback and other demographical factors helps us get accurate measures to improve the business.

The airline company previously made changes to improve Wi-Fi services to the passengers using conventional managerial insights , but there was not much increase in customer satisfaction.

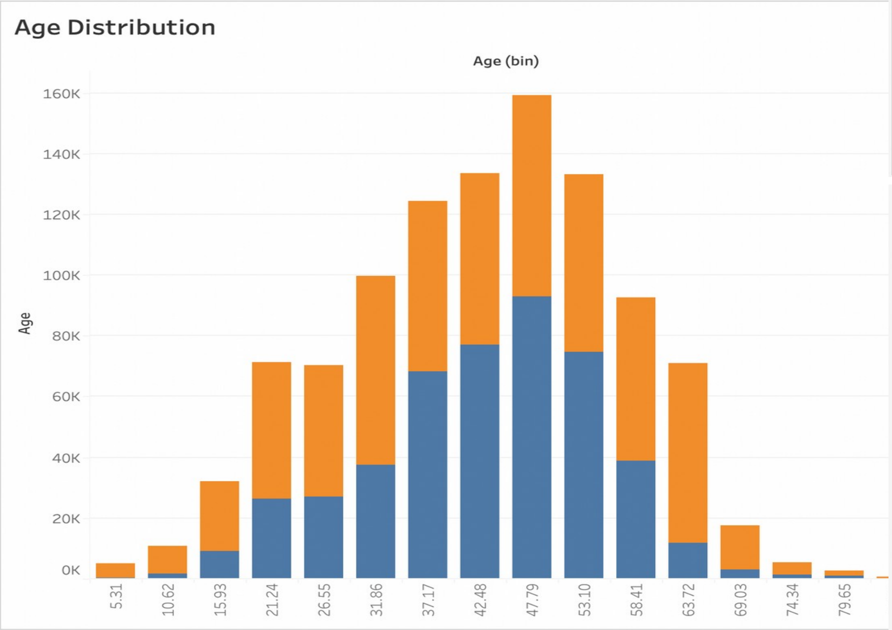
We are now using feedback data to find the facilities which can provide a considerable increase in passenger to improve the airline business.

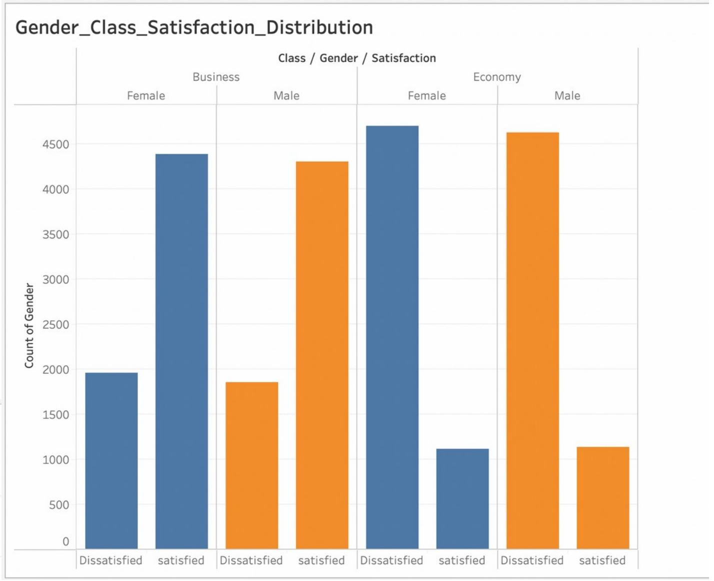
- We have ratings provided by around 100k customers for all the facilities offered by the airline, along with a final satisfied or neutral/dissatisfied feedback.

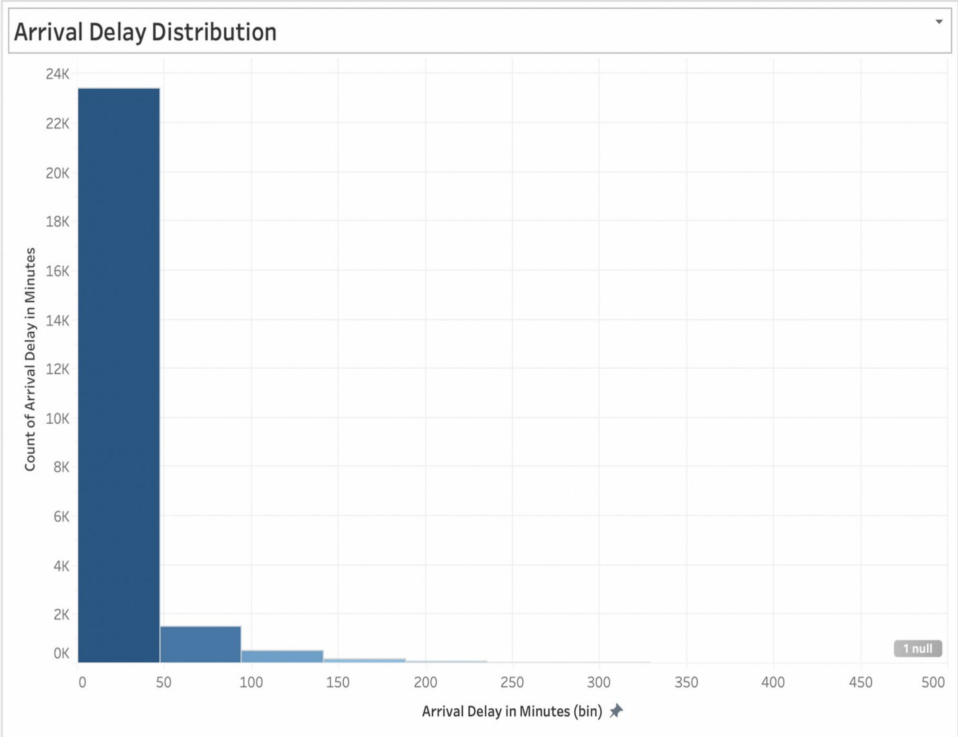
The dataset contains information about the passengers who travel on airlines - the column to predict is called satisfaction (TARGET VARIABLE).

1. Services that each passenger has signed up for -

- Type of Travel
- Class
- Flight Distance
- Inflight Wi-Fi service
- Departure
- Arrivals
- Baggage Claim
- Online-boarding
- Type of food
- Seat comfort
- Inflight entertainment
- Leg room service
- Check-in service

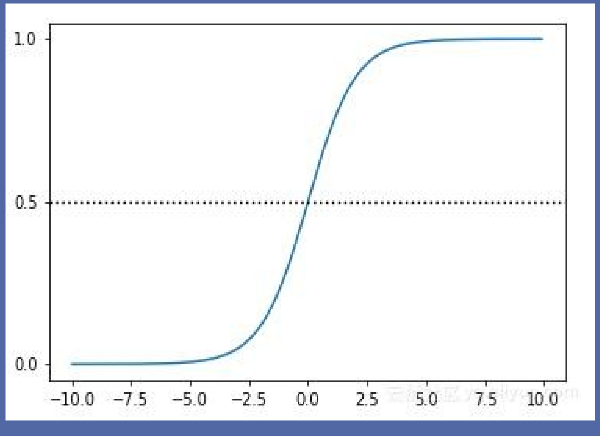

Age Distribution

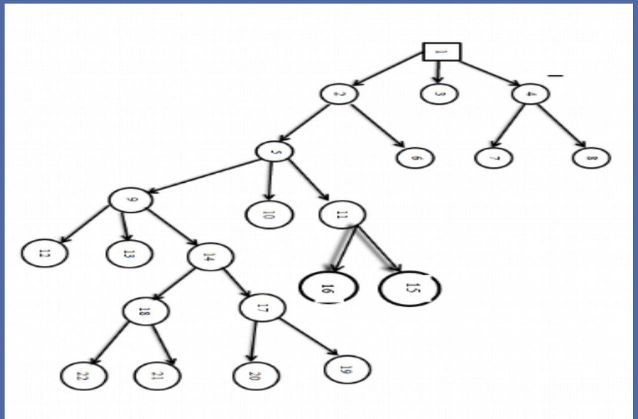

Gender_Class_Satisfaction_Distribution

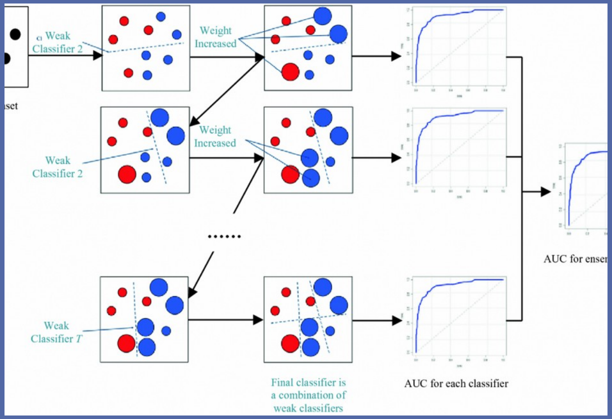

Arrival Delay Distribution

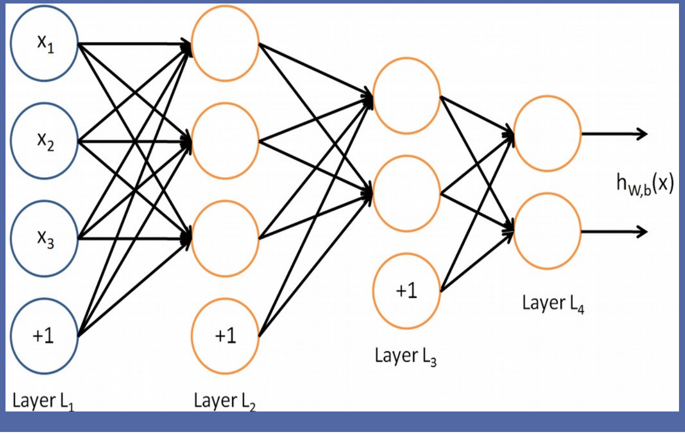
2. Demographic information about passengers –

- Id
- Gender
- Age Range
- Customer Type









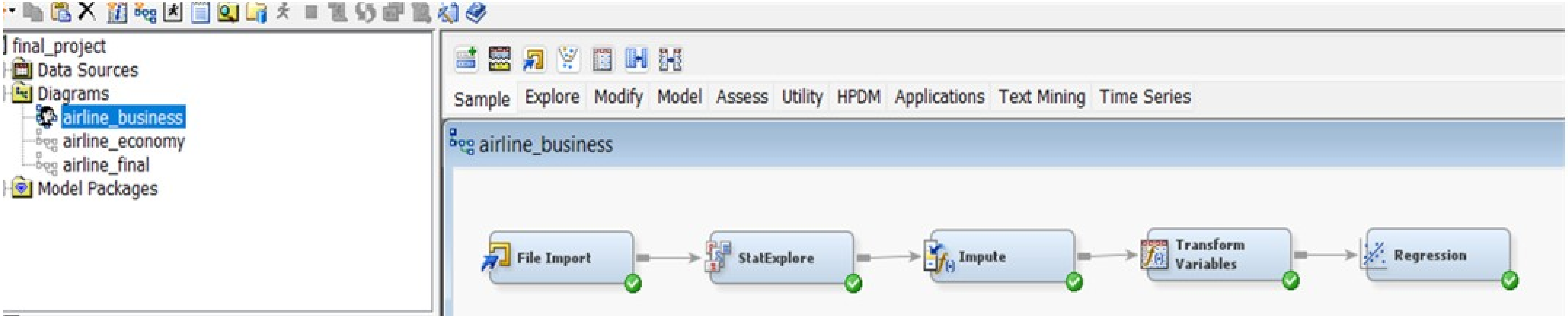
How we are using data to answer the business problem?

- We need to run regression models for business class and economy class separately.
- Get which variables are statistically significant for business class but have very less to none significance to economy class.
- Run the model for all observations with shortlisted statistically significantly variables of business class.

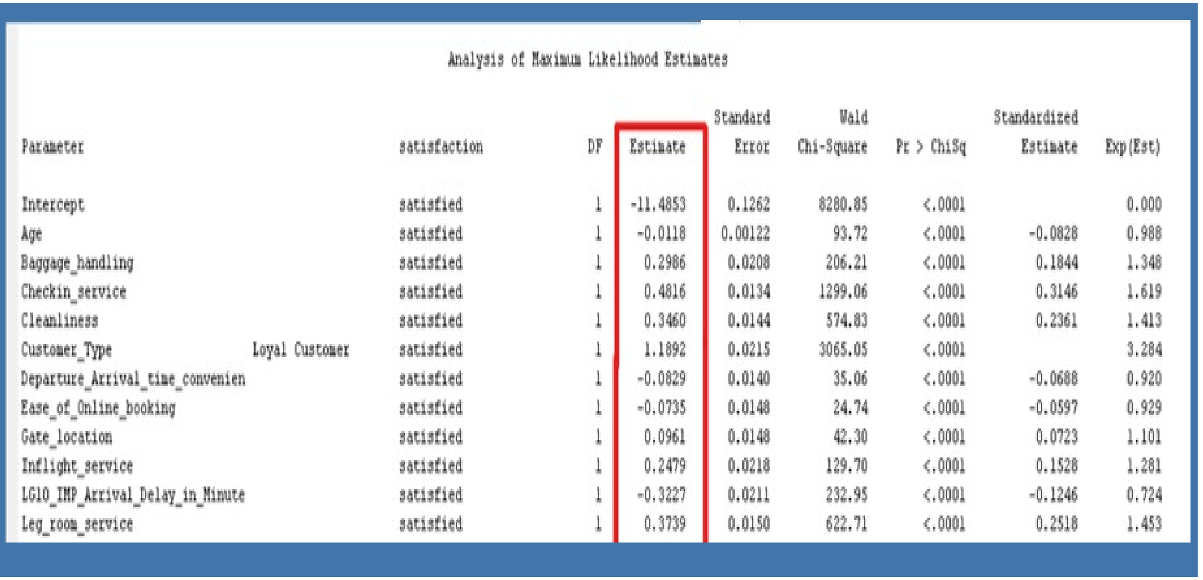
DATA MINING MODELS

- Outcome variable is binary – satisfied or neutral/dissatisfied.
- For finding the statistically significant variables we used logistic regression.
- For predicting satisfaction through statistically significant variables of business class, we used Decision tree, Gradient Boosting, Stepwise Logistic Regression and Neural Network.

Logistic Regression For Business and Economy Class Separately



- We have imported the file and explored the data.
- we have found that there are some missing values in the Arrival Delay in minutes. We have imputed the missing values with median values and transformed the skewed variables.
- After using logistic regression, we got variables that have a significant impact on the output (TARGET VARIABLE - Satisfaction).

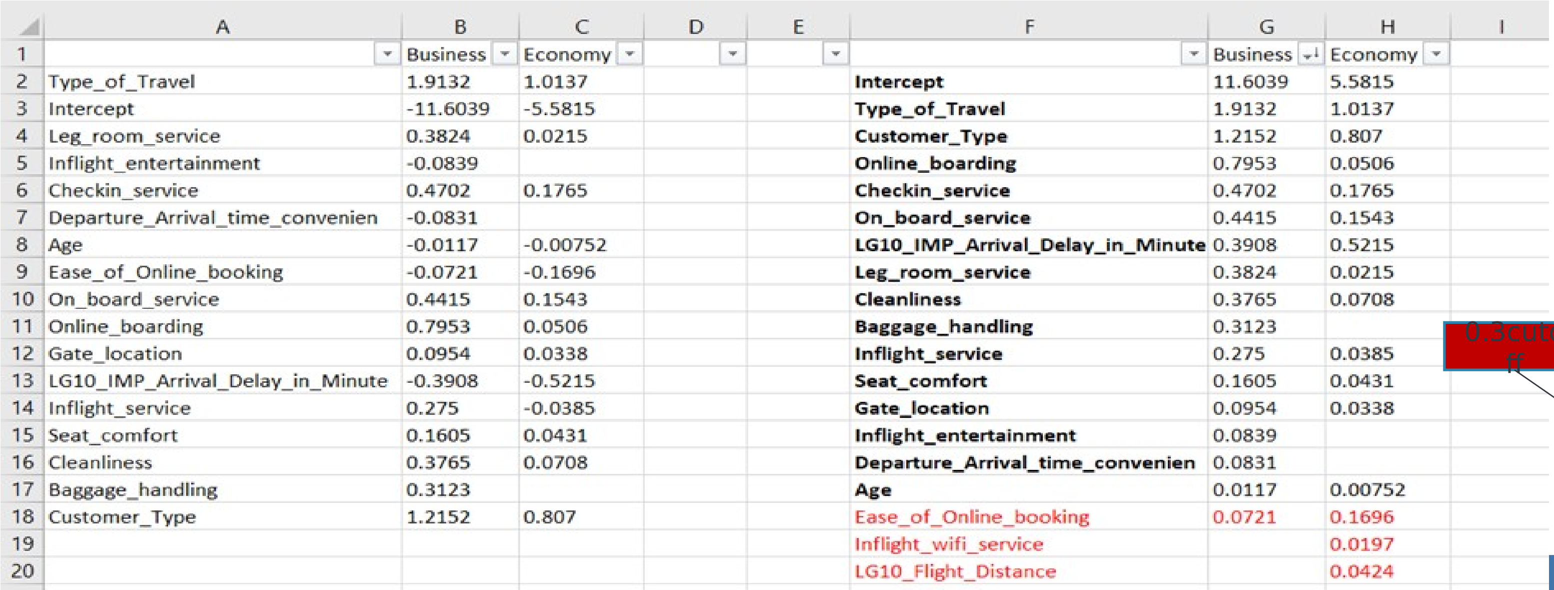


```
                              Analysis of Maximum Likelihood Estimates

                                                                    Standard      Wald                Standardized
Parameter                                    satisfaction   DF  Estimate     Error  Chi-Square  Pr > ChiSq    Estimate  Exp(Est)

Intercept                                    satisfied       1  -11.4853    0.1262     8280.85      <.0001                0.000
Age                                          satisfied       1   -0.0118   0.00122       93.72      <.0001     -0.0828    0.988
Baggage_handling                             satisfied       1    0.2986    0.0208      206.21      <.0001      0.1844    1.348
Checkin_service                              satisfied       1    0.4816    0.0134     1299.06      <.0001      0.3146    1.619
Cleanliness                                  satisfied       1    0.3460    0.0144      574.83      <.0001      0.2361    1.413
Customer_Type              Loyal Customer    satisfied       1    1.1892    0.0215     3065.05      <.0001                3.284
Departure_Arrival_time_convenien             satisfied       1   -0.0829    0.0140       35.06      <.0001     -0.0688    0.920
Ease_of_Online_booking                       satisfied       1   -0.0735    0.0148       24.74      <.0001     -0.0597    0.929
Gate_location                                satisfied       1    0.0961    0.0148       42.30      <.0001      0.0723    1.101
Inflight_service                             satisfied       1    0.2479    0.0218      129.70      <.0001      0.1528    1.281
LG10_IMP_Arrival_Delay_in_Minute             satisfied       1   -0.3227    0.0211      232.95      <.0001     -0.1246    0.724
Leg_room_service                             satisfied       1    0.3739    0.0150      622.71      <.0001      0.2518    1.453
```

Comparing the statistically Significant Variables



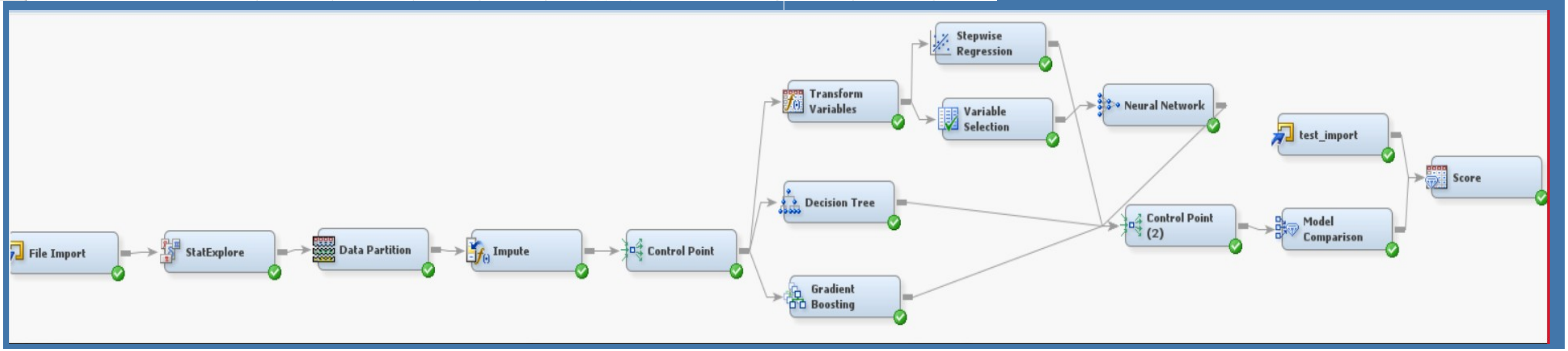
	A	B	C	D	E	F	G	H	I
1		Business	Economy				Business	Economy	
2	Type_of_Travel	1.9132	1.0137			**Intercept**	11.6039	5.5815	
3	Intercept	-11.6039	-5.5815			**Type_of_Travel**	1.9132	1.0137	
4	Leg_room_service	0.3824	0.0215			**Customer_Type**	1.2152	0.807	
5	Inflight_entertainment	-0.0839				**Online_boarding**	0.7953	0.0506	
6	Checkin_service	0.4702	0.1765			**Checkin_service**	0.4702	0.1765	
7	Departure_Arrival_time_convenien	-0.0831				**On_board_service**	0.4415	0.1543	
8	Age	-0.0117	-0.00752			**LG10_IMP_Arrival_Delay_in_Minute**	0.3908	0.5215	
9	Ease_of_Online_booking	-0.0721	-0.1696			**Leg_room_service**	0.3824	0.0215	
10	On_board_service	0.4415	0.1543			**Cleanliness**	0.3765	0.0708	
11	Online_boarding	0.7953	0.0506			**Baggage_handling**	0.3123		
12	Gate_location	0.0954	0.0338			**Inflight_service**	0.275	0.0385	
13	LG10_IMP_Arrival_Delay_in_Minute	-0.3908	-0.5215			**Seat_comfort**	0.1605	0.0431	
14	Inflight_service	0.275	-0.0385			**Gate_location**	0.0954	0.0338	
15	Seat_comfort	0.1605	0.0431			**Inflight_entertainment**	0.0839		
16	Cleanliness	0.3765	0.0708			**Departure_Arrival_time_convenien**	0.0831		
17	Baggage_handling	0.3123				**Age**	0.0117	0.00752	
18	Customer_Type	1.2152	0.807			Ease_of_Online_booking	0.0721	0.1696	
19						Inflight_wifi_service		0.0197	
20						LG10_Flight_Distance		0.0424	

0.3cutoff

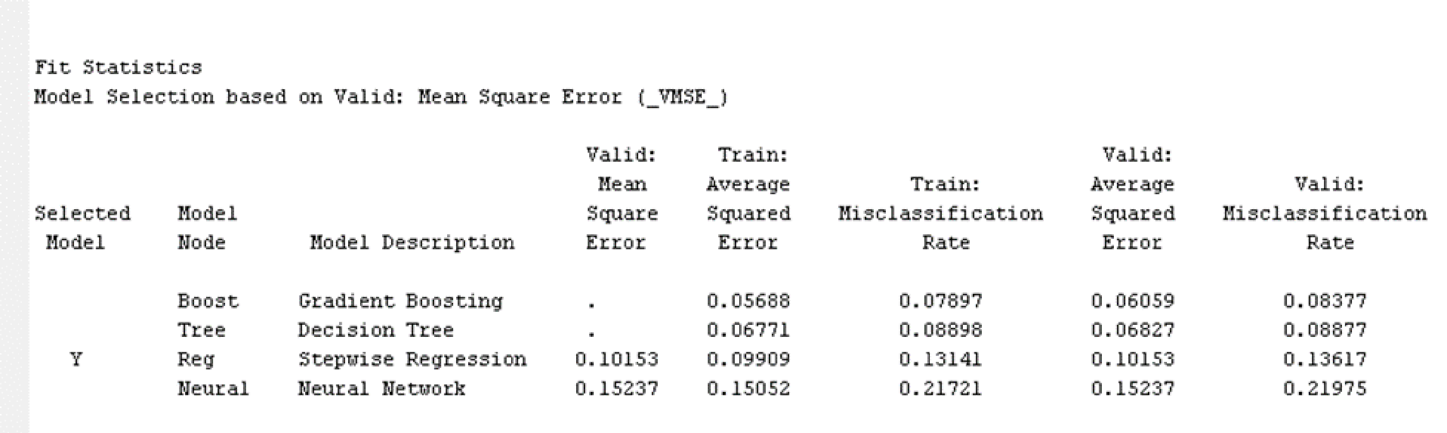
Exogenous business decision

- We have split the data into 80 percent training data and 20percent validation data to get the statistical distribution of variables.

MODEL	ACCURACY %
Stepwise logistic regression	86.3831
Decision trees with entropy criterion	91.1325
Gradient boosting	91.623
Deep learning model with neural network	78.0253



Model Comparison, Interpretation and Scoring

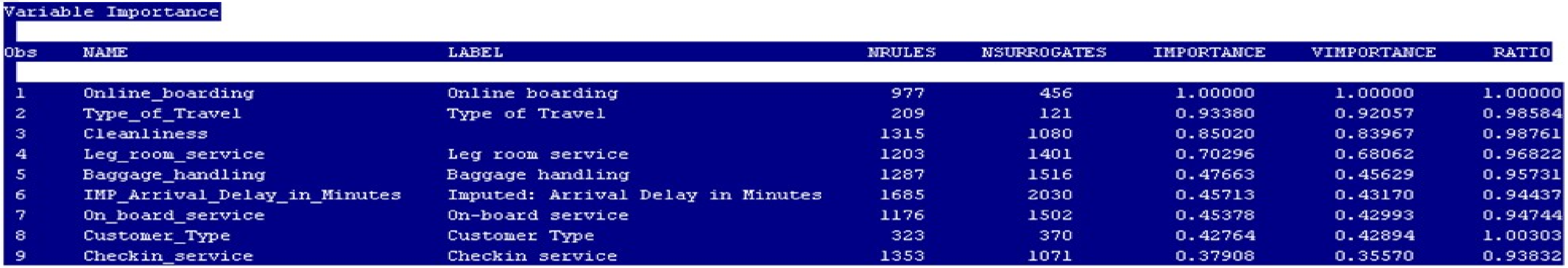


```
Fit Statistics
Model Selection based on Valid: Mean Square Error (_VMSE_)

                                                   Valid:     Train:                          Valid:
                                                    Mean      Average        Train:           Average          Valid:
Selected    Model                                  Square     Squared    Misclassification    Squared    Misclassification
 Model      Node        Model Description          Error      Error            Rate           Error            Rate

            Boost      Gradient Boosting             .        0.05688        0.07897          0.06059        0.08377
            Tree       Decision Tree                 .        0.06771        0.08898          0.06827        0.08877
   Y        Reg        Stepwise Regression         0.10153    0.09909        0.13141          0.10153        0.13617
            Neural     Neural Network              0.15237    0.15052        0.21721          0.15237        0.21975
```

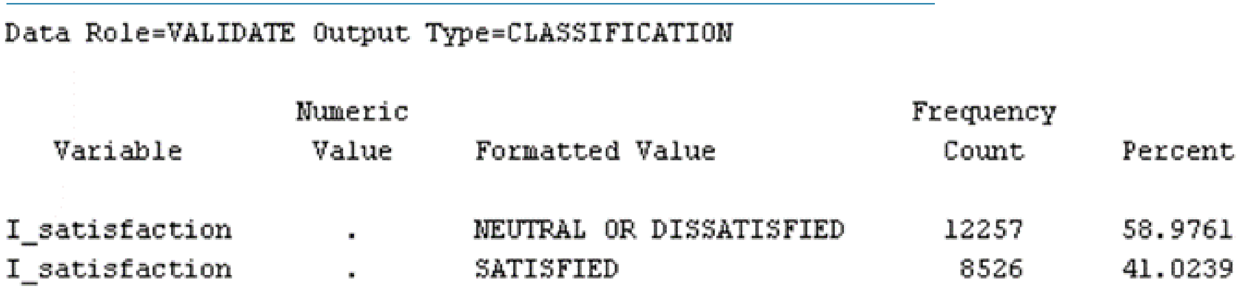
- After comparing all the models with test data, we have got the lowest MSE for Gradient Boosting and it also has the lowest MIS rate.
- Misclassification rate = **FP+FN/All values.**



```
Variable Importance

Obs    NAME                          LABEL                                 NRULES    NSURROGATES    IMPORTANCE    VIMPORTANCE    RATIO

 1     Online_boarding               Online boarding                          977            456       1.00000        1.00000   1.00000
 2     Type_of_Travel                Type of Travel                           209            121       0.93380        0.92057   0.98584
 3     Cleanliness                                                           1315           1080       0.85020        0.83967   0.98761
 4     Leg_room_service              Leg room service                        1203           1401       0.70296        0.68062   0.96822
 5     Baggage_handling              Baggage handling                        1287           1516       0.47663        0.45629   0.95731
 6     IMP_Arrival_Delay_in_Minutes  Imputed: Arrival Delay in Minutes       1685           2030       0.45713        0.43170   0.94437
 7     On_board_service              On-board service                        1176           1502       0.45378        0.42993   0.94744
 8     Customer_Type                 Customer Type                            323            370       0.42764        0.42894   1.00303
 9     Checkin_service               Checkin service                         1353           1071       0.37908        0.35570   0.93832
```



```
Data Role=VALIDATE Output Type=CLASSIFICATION

                    Numeric                                     Frequency
  Variable           Value      Formatted Value                   Count        Percent

I_satisfaction         .        NEUTRAL OR DISSATISFIED           12257        58.9761
I_satisfaction         .        SATISFIED                          8526        41.0239
```

- We have scored the new data which is not yet exposed to model.
- The final scoring with new data has given us the distribution of satisfaction and dissatisfaction as

Conclusion

More significant variables, put more revenue.

From our results, we see the below variable gives significant change in their business class revenue. Online_Boarding, Type of travel, Cleanliness, Leg_room_service, Baggage Handling, IMP_Arrival_Delay_In_Minutes, Onboard_Service,Customer_Type, Checkin_Service.

We see the variables Online_Boarding, Type of travel, Cleanliness, Leg_room_service gives major impact on improving the business class passengers. We suggest managerial team to make changes to these variables to see the significant change in passengers.

We suggest management to make changes such as ease of access to Online_boarding, taking extra care in cleaning the business class area, targeting the people who are travelling for business purposes and making changes in the Leg_room_service.