Capstone Project - 2

Team 2

Taxi Mobility Surge Price Prediction

Team Members

Bhanu Pratap Shahi
Sanchita Paul
Shubham Kumar
Sumanta Muduli

Content

Title

- Problem Statement
- Data Summary
- Comparing features with Surge Pricing Type
- Feature Selection
- Models used
- Which model did we choose and why?
- Challenges
- Conclusion

Problem Statement

OP

- The goal is to build a predictive model which can help Sigma Cabs in predicting Surge Pricing Type proactively.
- This will help them in matching the right priced cabs with the right customers quickly and efficiently.

Data Summary:

Data set name Data Sigma Cabs

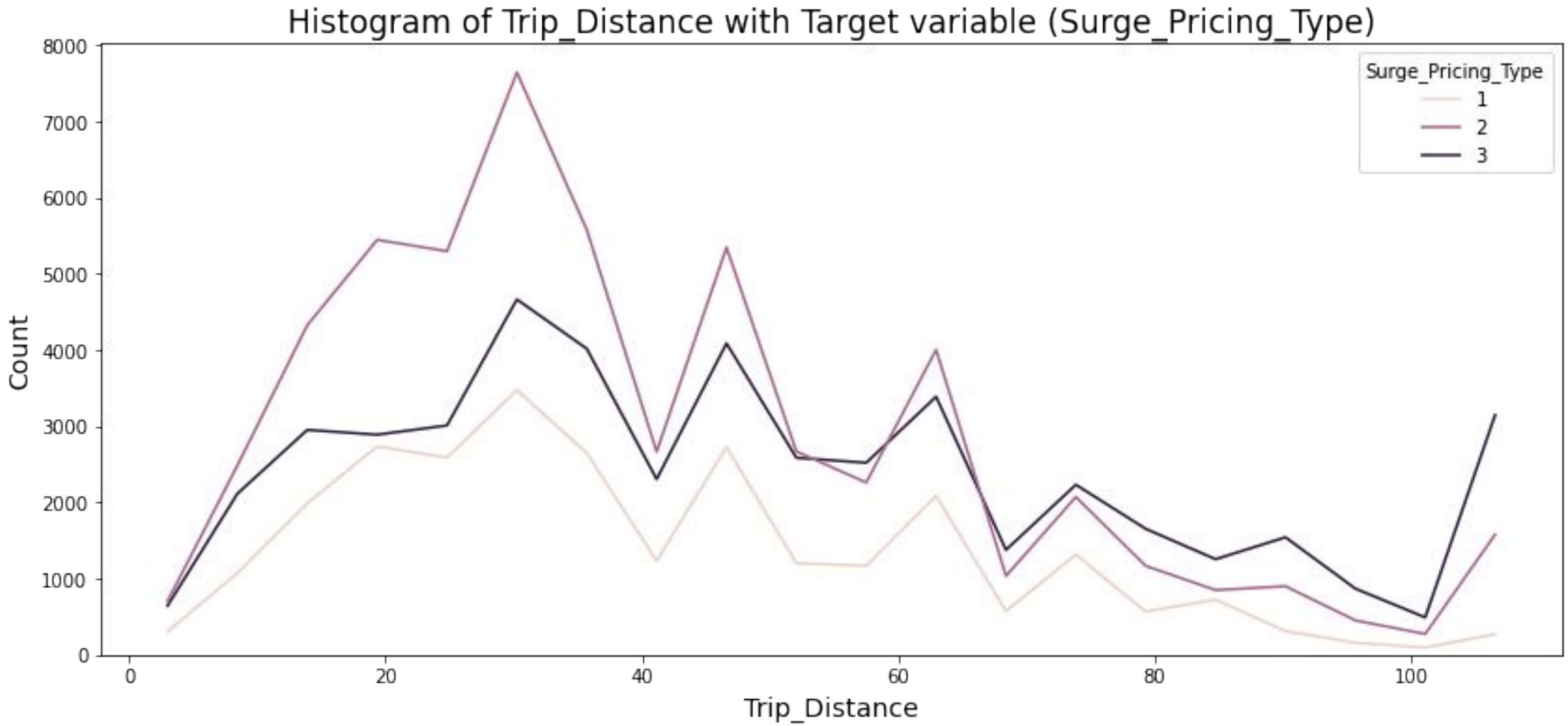
Shape

- Rows -- 131,662
- Columns--14

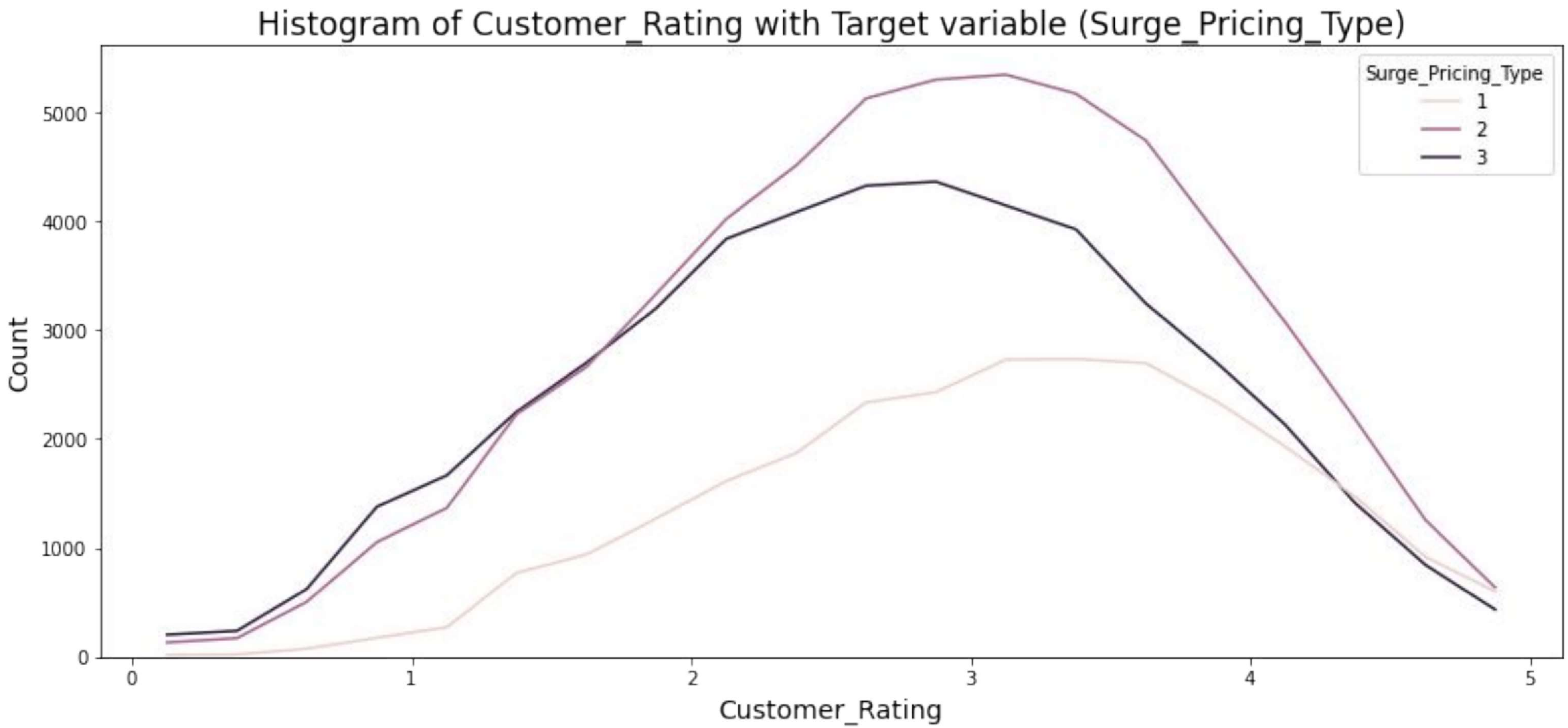
Features

Trip_ID,Trip_Distance,Type_of_Cab,Customer_since_months, Life_Style_Index, Confidence_Life_Style_Index,Destination_Type, Customer_Rating,Cancellation_Last_1Month,Var1,Var2,Var3,Gender, Surge_Pricing_Type

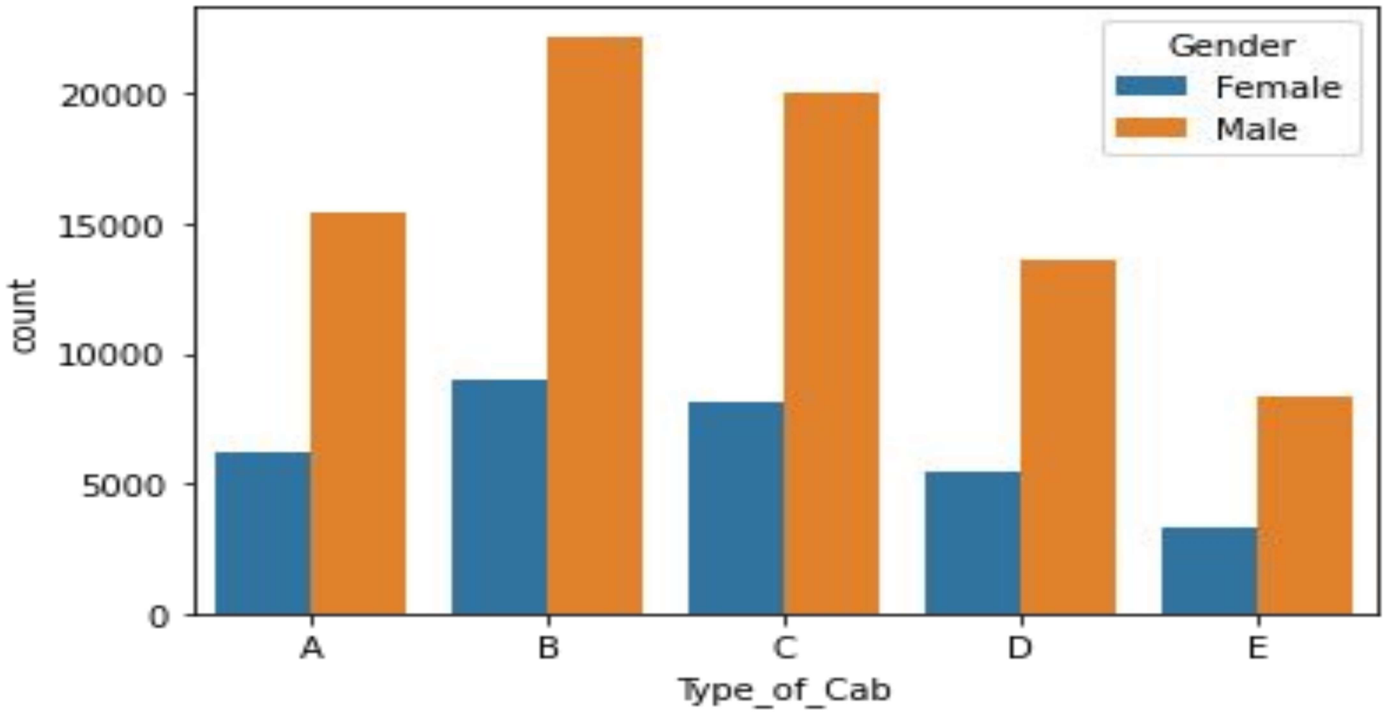
Comparing Trip Distance with Surge Pricing Type:



Comparing Customer Rating with Surge Pricing Type:



Count of Type of Cab with Gender Filter

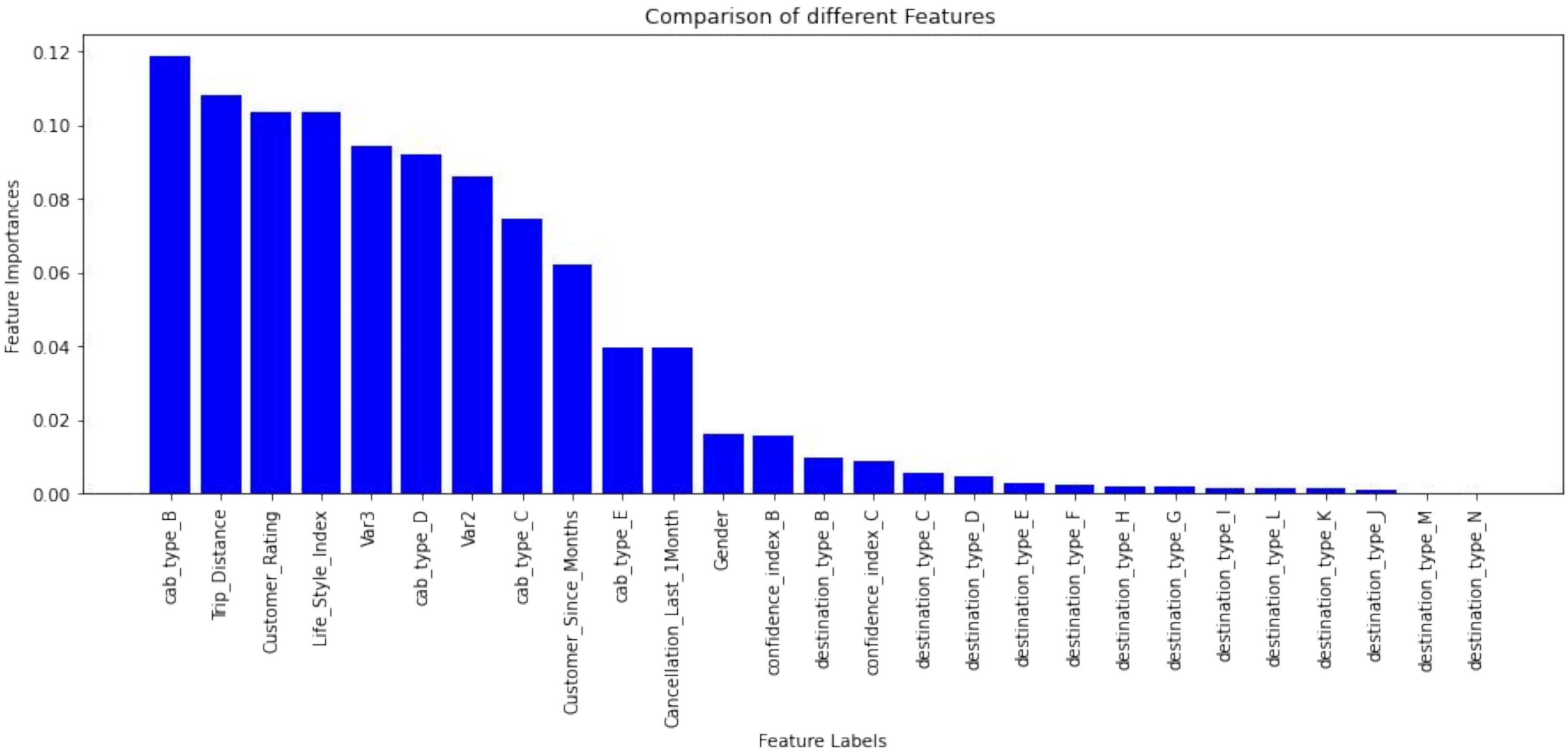


Features selection:

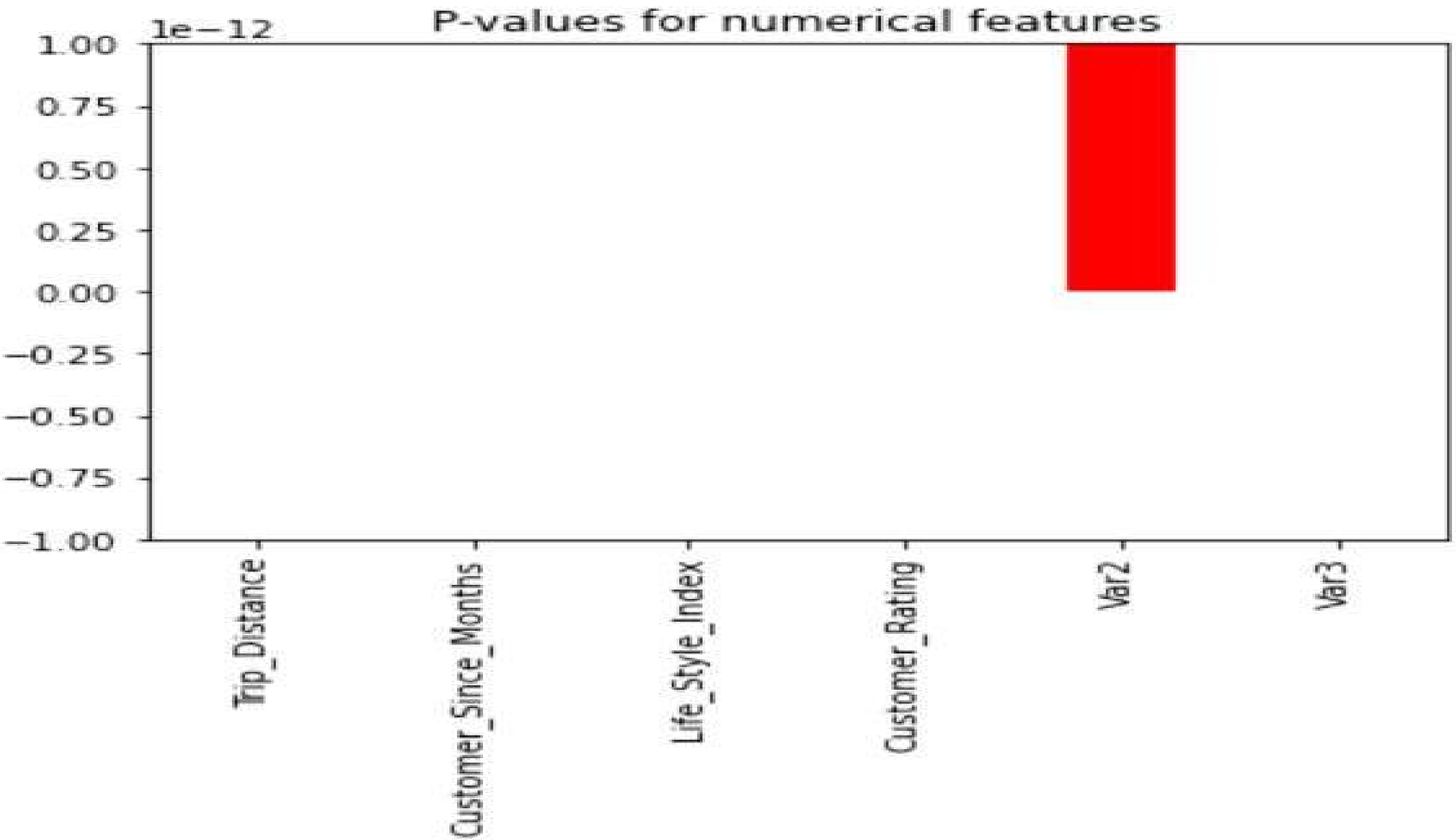
Methods used:

- **Extra Tree Classifier**
- **ANOVA**
- **Chi-Square**

Extra Trees Classifier:

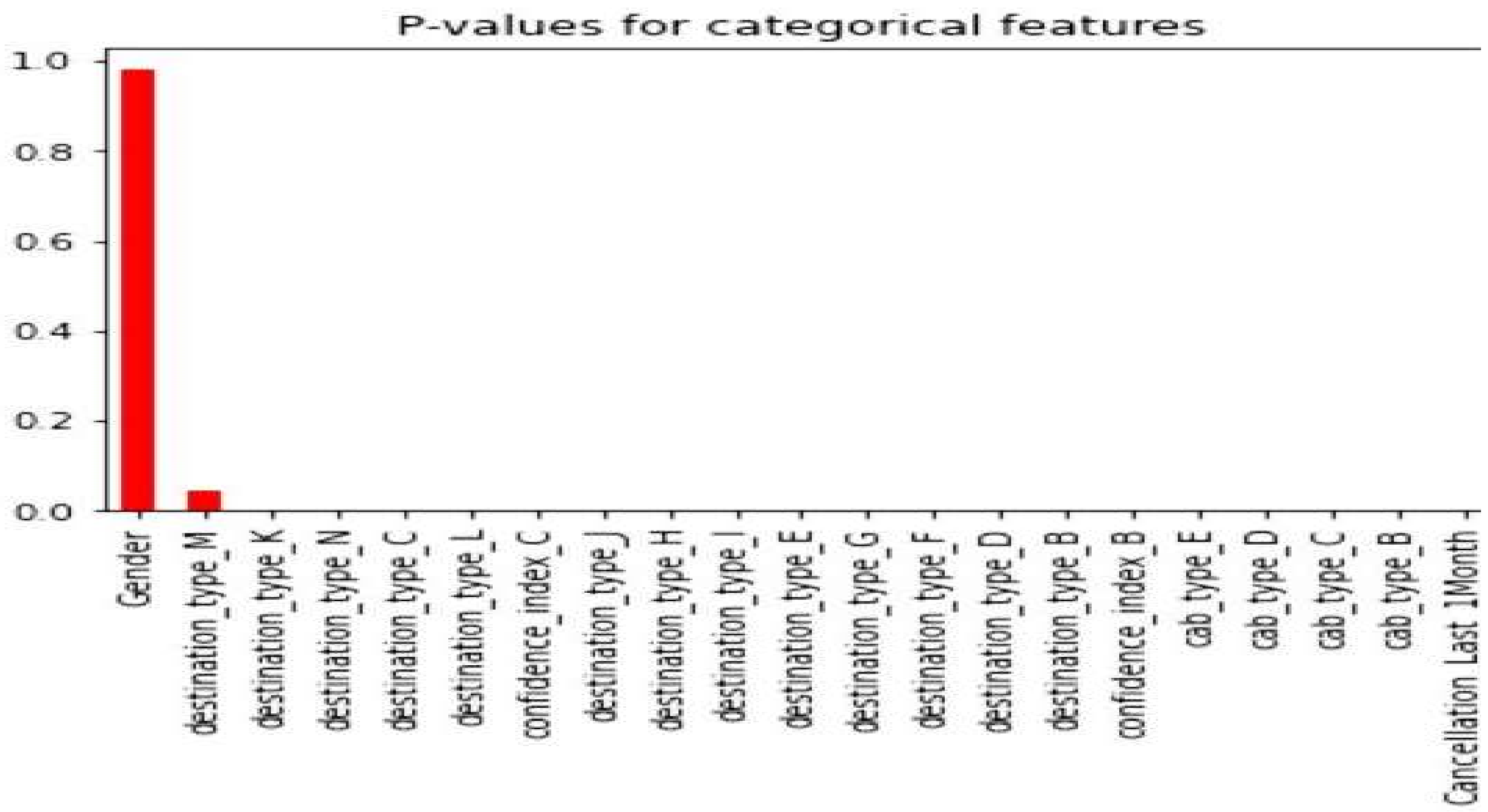


ANOVA:



P-values for numerical features
1e−12
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Trip_Distance
Customer_Since_Months
Life_Style_Index
Customer_Rating
Var2
Var3

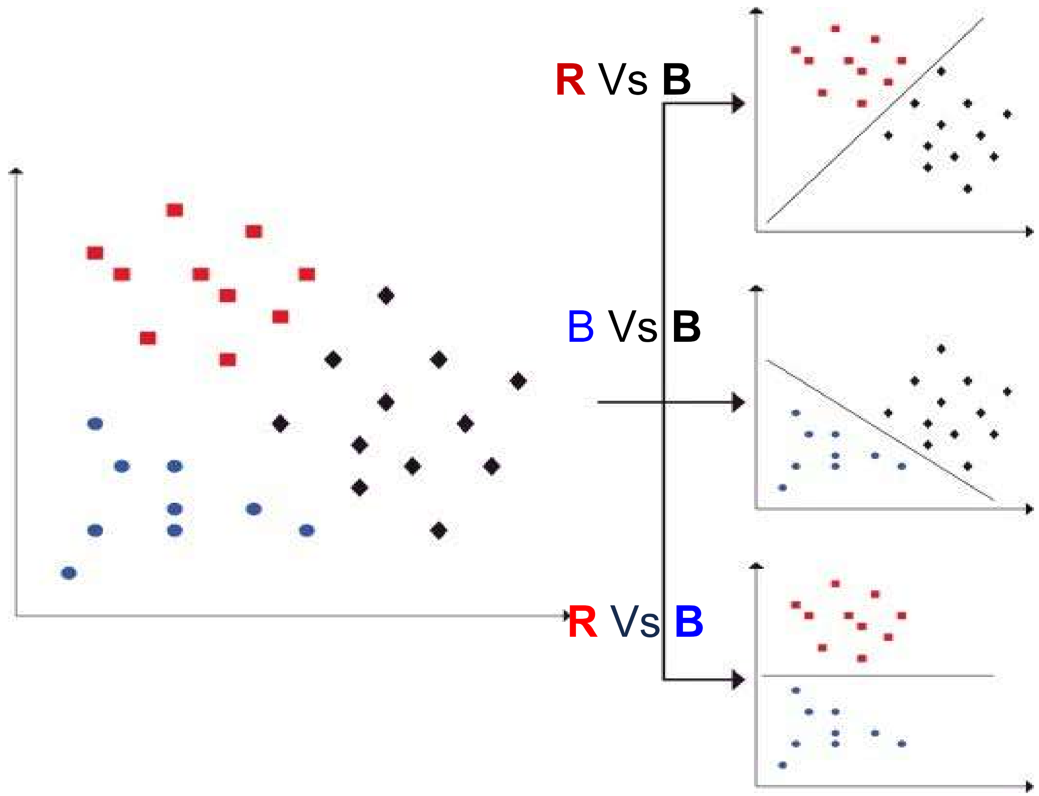
Chi-Square:


P-values for categorical features
1.0
0.8
0.6
0.4
0.2
0.0
Gender
destination_type_M
destination_type_K
destination_type_N
destination_type_C
destination_type_L
confidence_index_C
destination_type_J
destination_type_H
destination_type_I
destination_type_E
destination_type_G
destination_type_F
destination_type_D
destination_type_B
confidence_index_B
cab_type_E
cab_type_D
cab_type_C
cab_type_B
Cancellation_Last_1Month

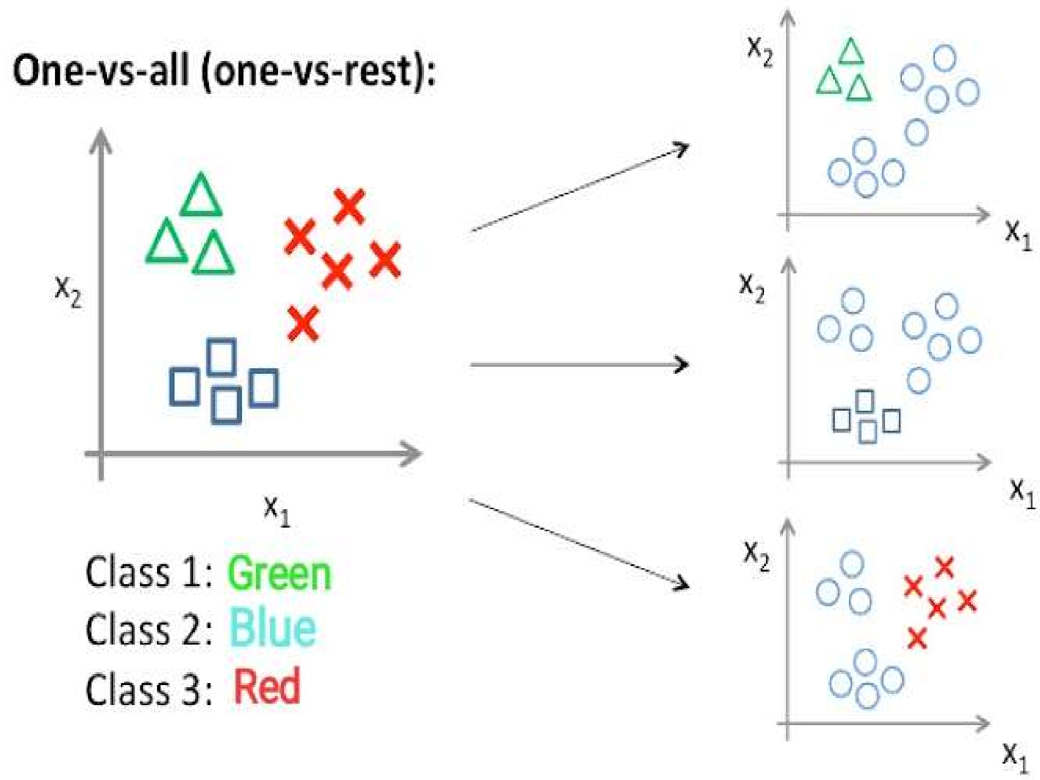
Models used:

- Logistic Regression Classifier
- SVM Classifier
- Random Forest Classifier
- XGBoost Classifier

One vs One and One vs Rest:



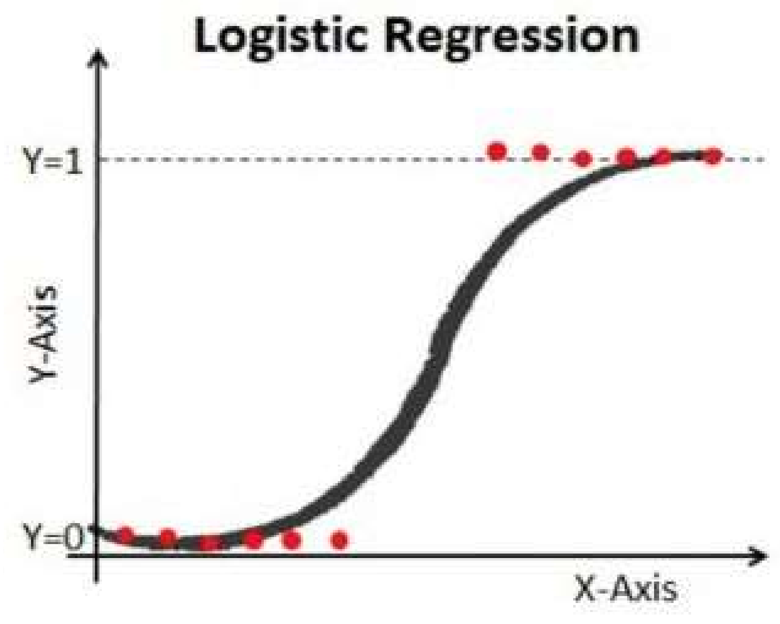
One vs One



One vs Rest

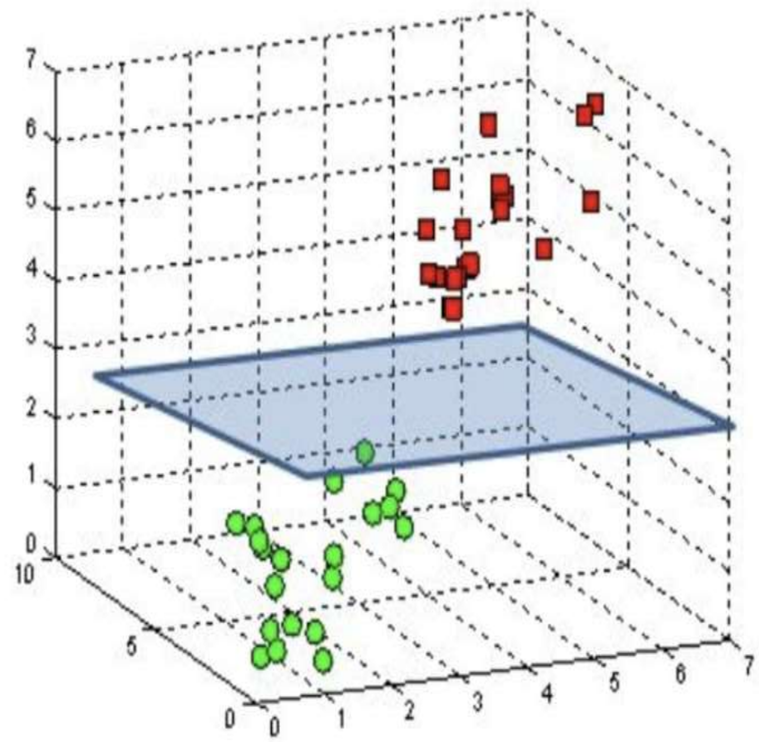
Logistic Regression:

- One vs Rest approach (“ovr”)
- Hyperparameter Tuning(Bayesian Optimisation)-C:0.001, solver:”lbfgs”,penalty=l2
- Metric Scores- Accuracy=72%, Precision=72%, Recall=70% & f1_score=71%



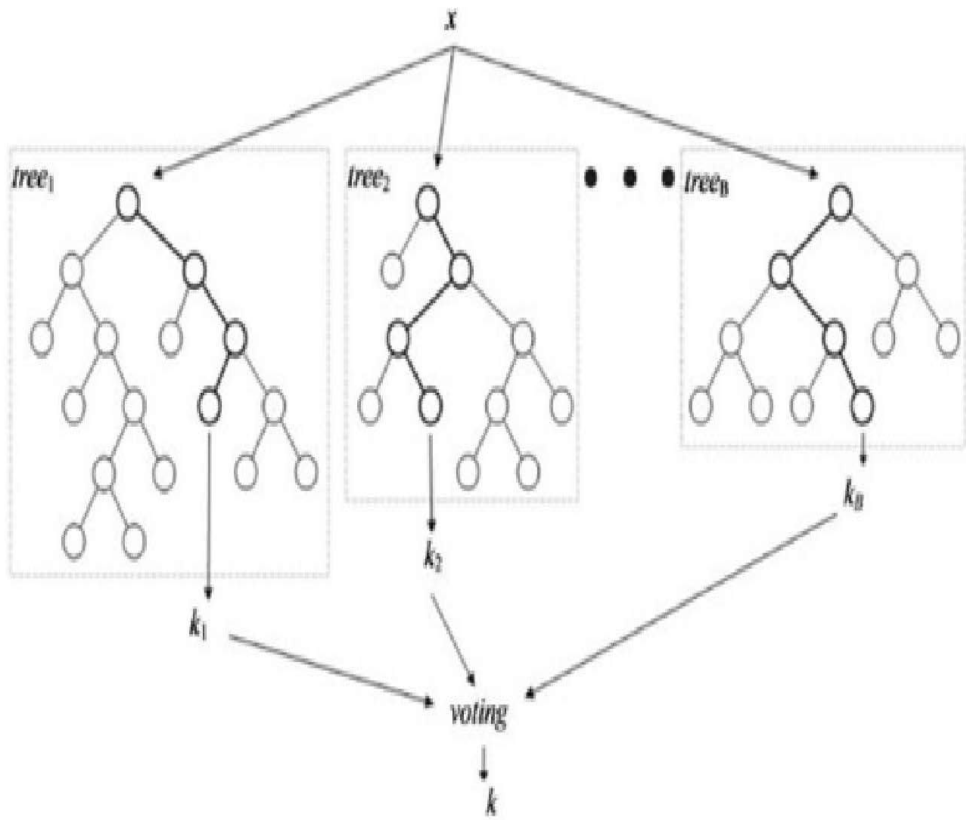
Support Vector Machine:

- One vs One approach (“ovo”)
- Parameters - C:1, degree =3,
- Kernel - Poly Kernel is giving us the best results. Accuracy i.e 72%, Precision=73%, Recall=70% & f1_score=70%



Random Forest Classifier:

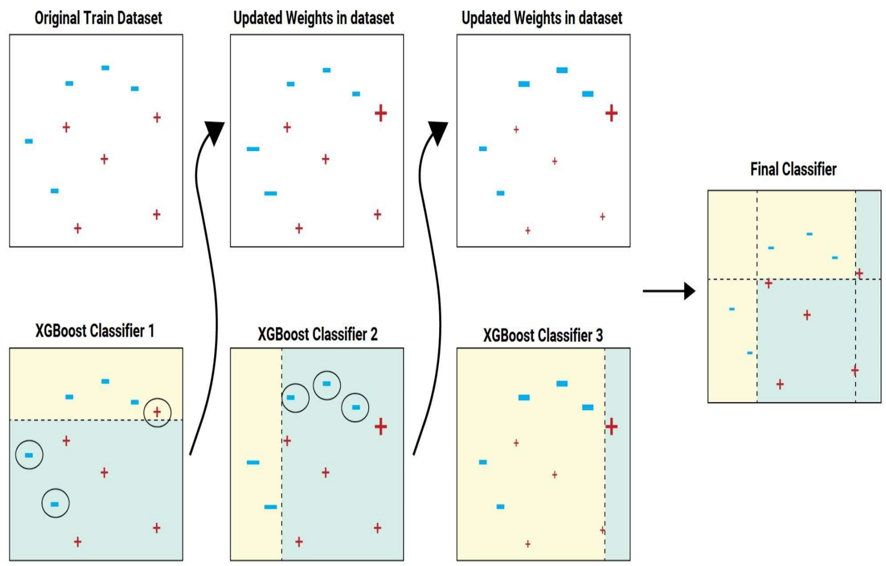
- Hyper parameter Tuning(Bayesian Search)-
  ('max_depth', 8),
  ('min_samples_leaf', 10),
  ('min_samples_split', 50),
  ('n_estimators', 100)
- accuracy= 72%,
  precision=73%,
  recall=70%,fi_score=71%



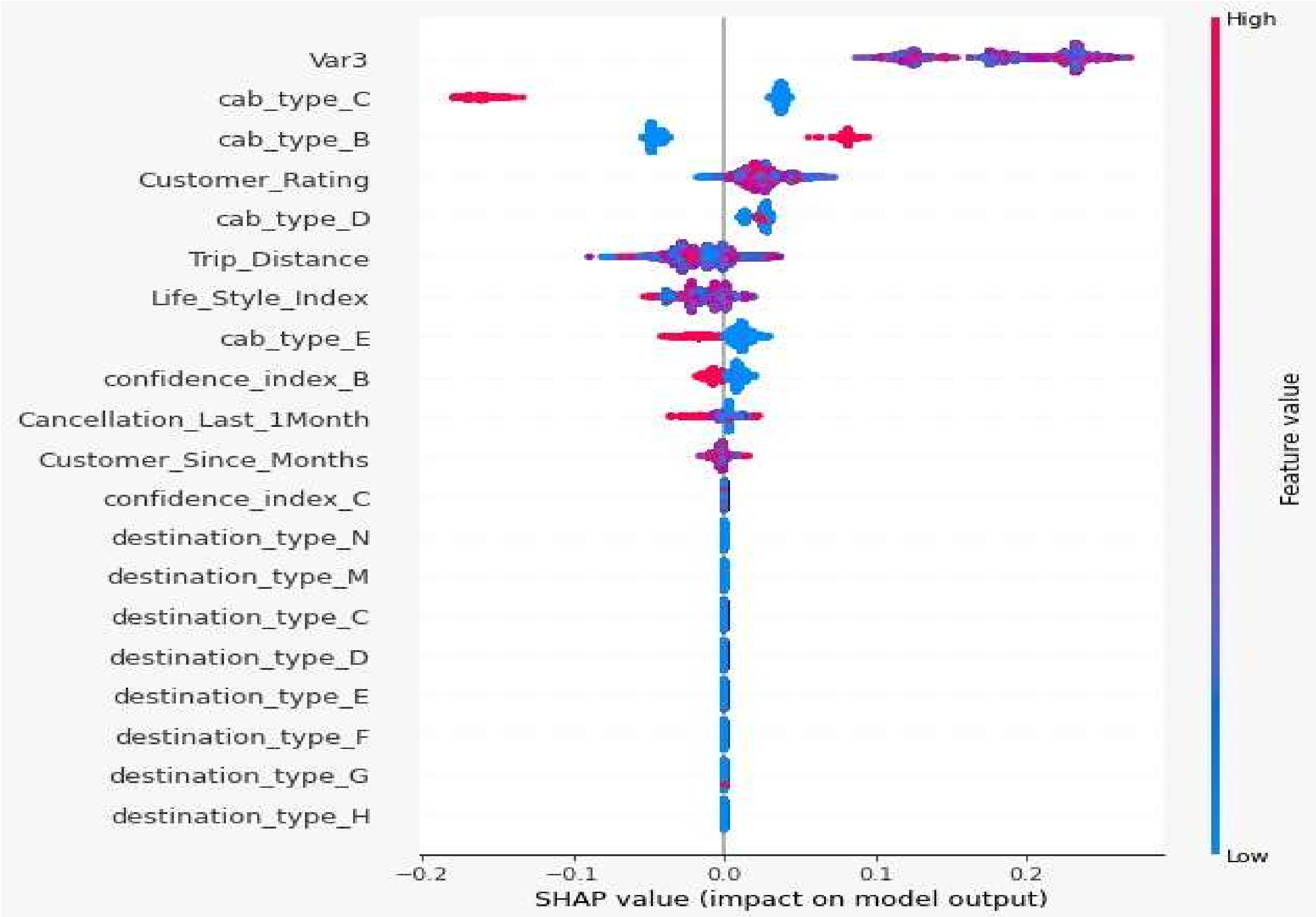
XGBoost Classifier:

- Hyperparameters-gamma=0, learning_rate=0.1, max_depth=15, n_estimators=100, objective='multi:softprob'

- Metric Scores- accuracy=72%, precision=73%, recall=70%,f1_score=71%



SHAP Values:



Var3
cab_type_C
cab_type_B
Customer_Rating
cab_type_D
Trip_Distance
Life_Style_Index
cab_type_E
confidence_index_B
Cancellation_Last_1Month
Customer_Since_Months
confidence_index_C
destination_type_N
destination_type_M
destination_type_C
destination_type_D
destination_type_E
destination_type_F
destination_type_G
destination_type_H
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Low
Feature value

Which model did we choose and why?

- We choose logistic regression as it’s evaluation scores is very similar to other complicated models but it is computationally cheaper and more interpretable.
- Accuracy : 72%
- Recall : 72%
- Precision : 72%
- This is the most consistent performing model with same scores for all metrics.

Challenges

- Lots of NaN values in the dataset.
- Some features like Var1,Var2,Var3 are not clearly explained.
- Choosing the right encoding technique for categorical features.
- Choosing the right features for modelling.
- Faced issues while running the models as the dataset is large.
- Choosing the right models as there is not much difference in accuracy.

AI

Conclusion

- We build a predictive model which can help Sigma Cabs in predicting Surge Pricing Types proactively.
- This will helps in matching the right cab with the right customer quickly and efficiently
- They can increase their customer base and profit by providing better services.

Q & A