

**PGP DSE**

**CAPSTONE PROJECT**

FINAL REPORT

**TRIP PRICING WITH TAXI MOBILITY ANALYTICS**

**Done by:**

Vignesh. S

Abhijith S Varma

Siva Kumar. G

Ram Prakash. V

Vignesh .M

**Mentor:**

Ms. Vibha Santhanam

**TABLE OF CONTENTS:**

|  |  |  |
| --- | --- | --- |
| **SI NO** | **TITLE** | **PAGE**  **NUMBER** |
| 1 | BUSINESS UNDERSTANDING | 2 |
| 2 | DATA AND FINDINGS | 2 |
| 3 | OVERVIEW OF FINAL PROCESS | 4 |
| 4 | STEP BY STEP WALK THROUGH THE SOLUTION | 5 |
| 5 | MODEL EVALUATION | 5 |
| 6 | COMPARISION TO BENCH MARK | 6 |
| 7 | VISUALIZATION/BUSINESS INSIGHTS | 7 |
| 8 | IMPLICATIONS | 14 |
| 9 | LIMITATIONS | 14 |
| 10 | CLOSING REFLECTIONS/FUTURE SCOPE | 14 |

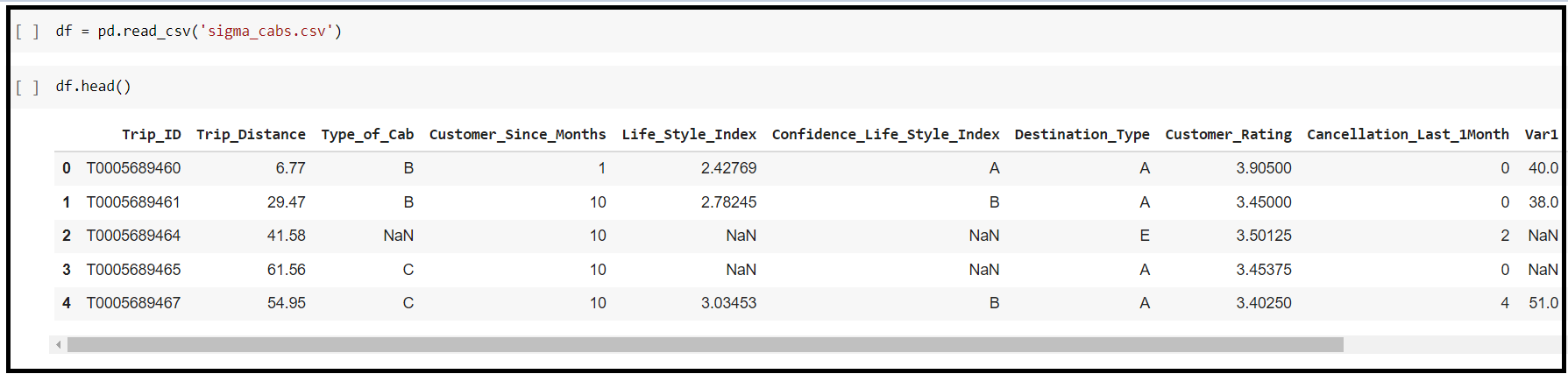
**BUSINESS UNDERSTANDING:**

We are predicting the surge price type for Sigma Cabs. Previously surge price was given by service providers, from that information they have captured surge price type, we are building a predictive model based on that surge price type, so that they can fix the fare beforehand.

**DATA AND FINDINGS:**

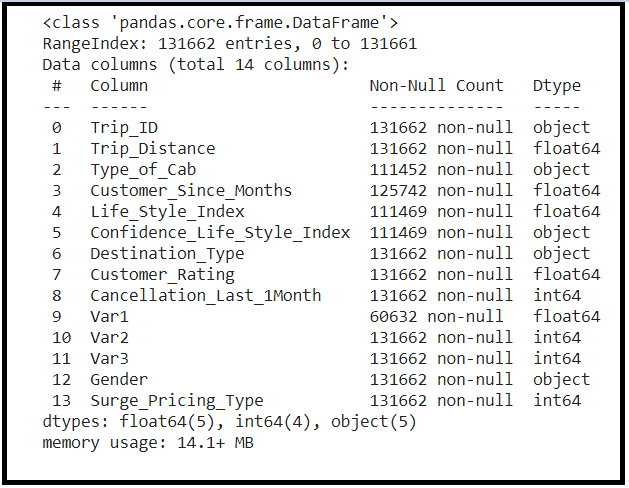
1. **DATA:**

STEP 1: Read the dataset.



STEP 2: Understand the dataset by checking

* Info
* Nunique
* Describe
* Shape



**VARIABLE CATEGORIZATION:**

* No of rows: 131662
* Total Features/Columns :14
* Numerical Features :06
* Categorical Features:08
* Target: Surge\_Pricing\_Type (Multiclassification)

1. **FINDINGS:**

* F1- weighted score improved after changing the null value imputation method

Before the new null value imputation method:

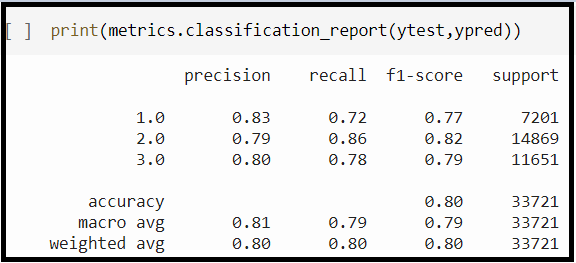
**F1-Weighted-Score =0.69**

After the new null value imputation method:

**F1-Weighted-Score =0.80**

* Model Building with smote and without smote

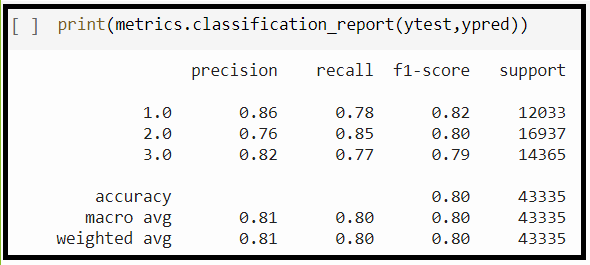
**BEFORE SMOTE:**



**INFERENCE:**

Scores of class1 was generally inconsistent in comparison with other classes.

**AFTER SMOTE:**



**INFERENCE:**

Scores of class 1 became inline with scores of other classes

**SCALING:**

* Model building with Standard scaler and Min Max scaler.

**INFERENCE:**

There is not much difference in the performance of the models.

So, we took data with Min max scaling for further model building.

**OVERVIEW OF THE FINAL PROCESS:**

**SALIENT FEATURES:**

* This a Multi-classification dataset.
* This is a commercial dataset based on taxi surge type
* There is some confidentiality involved in this data such as Masked variables and Masked target.

**Data Preprocessing**

* Data type conversion
* Missing value Treatment
* Outlier Treatment
* Feature selection using Multicollinearity
* Scaling
* Encoding using dummies
* Treating class imbalance

**Modelling:**

* Base Models can be built using KNN.
* Furthermore, we and did one standalone model (Decision Tree) and three Ensemble models (Random Forest, Adaboost and Xgboost) using cross-validation method.
* After that we did stacking model using three of the best models from the above (Random Forest, Adaboost and Xgboost) and bagging model of that particular stacking model.
* Finally, we did Grid search hyper parameter tuning on the best model (stacking model mentioned above).

**Evaluation:**

* We use cross-validation method to obtain the average f1-weighted score.
* From this cross-validation method, we evaluate model by analyzing the following metrics:

**F1-score (Weighted Fi-score)**

* Finally, we check whether the model is under fitting or overfitting by evaluating the Bias and Variance error.
* If the Bias and variance error in optimal level we draw inferences from the model.

**Deployment:**

* We are deploying the model as a web application using Flask, HTML5 and CSS.

**STEP BY STEP WALKTHROUGH OF THE SOLUTION:**

**STEP 1: Read the dataset**

**STEP 2: Understand the dataset by checking**

* Info
* Nunique
* Describe
* Shape

**STEP 3: Data Preprocessing**

* Data type conversion
* Missing value Treatment
* Outlier Treatment
* Feature selection using Multicollinearity
* Scaling
* Encoding using dummies
* Treating class imbalance

**STEP4: MODEL BUILDING**

* **Base Models** can be built using KNN.
* Furthermore, we and did one standalone model (Decision Tree) and three Ensemble models (Random Forest, Adaboost and Xgboost) using cross-validation method.
* After that we did stacking model using three of the best models from the above (Random Forest, Adaboost and Xgboost) and bagging model of that particular stacking model.
* Finally, we did Grid search hyper parameter tuning on the best model (Random Forest model mentioned above).

**STEP 5: MODEL DEPLOYMENT**

* We are deploying the model as a web application using Flask, HTML5 and CSS

**MODEL EVALUATION:**

* We use cross-validation method to obtain the average f1-weighted score.
* From this cross-validation method, we evaluate model by analyzing the following metrics:

**F1-score (Weighted Fi-score)**

* Finally, we check whether the model is under fitting or overfitting by evaluating the Bias and Variance error.
* If the Bias and variance error in optimal level we draw inferences from the model.

**Final model-Random Forest:**

* The best model was stacking model but as the stacking model and bagging model is based on Random Forest.
* we are proceeding with hyper parameter tuning for Random Model.
* Random Forest is an ensemble modelling technique which uses bagging of multiple Decision trees and majority of the individual decision tree’s prediction will be taken as the prediction for random forest.

**Objective:**

* Objective is to build a predictive model on the surge pricing type.

**Parameters:**

These are all the parameters for the final model

* *n\_estimators*
* *criterion*
* *max\_depth*
* *scoring*
* *cv*

**We Evaluate the success by F1-weighted score.**

**We improved the robustness of the model by balancing the class weights of the target variables which is surge pricing type.**

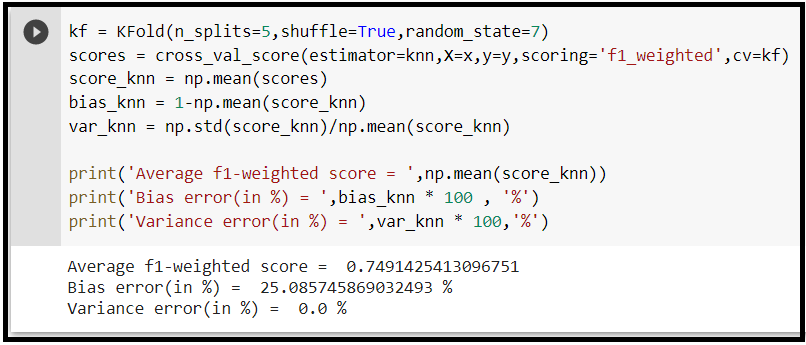
**COMPARISION TO BENCHMARK:**

**KNN model is the Base model**

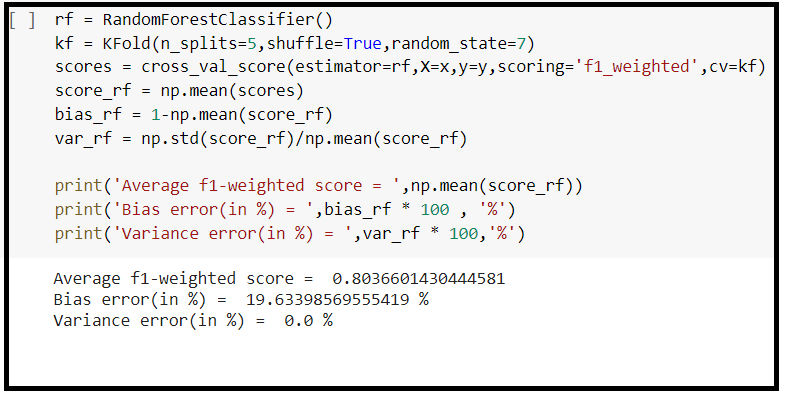
**Random Forest is the Final Model.**



**BASE MODEL-K NEAREST NEIGHBOUR**



**FINAL MODEL -RANDOM FOREST:**

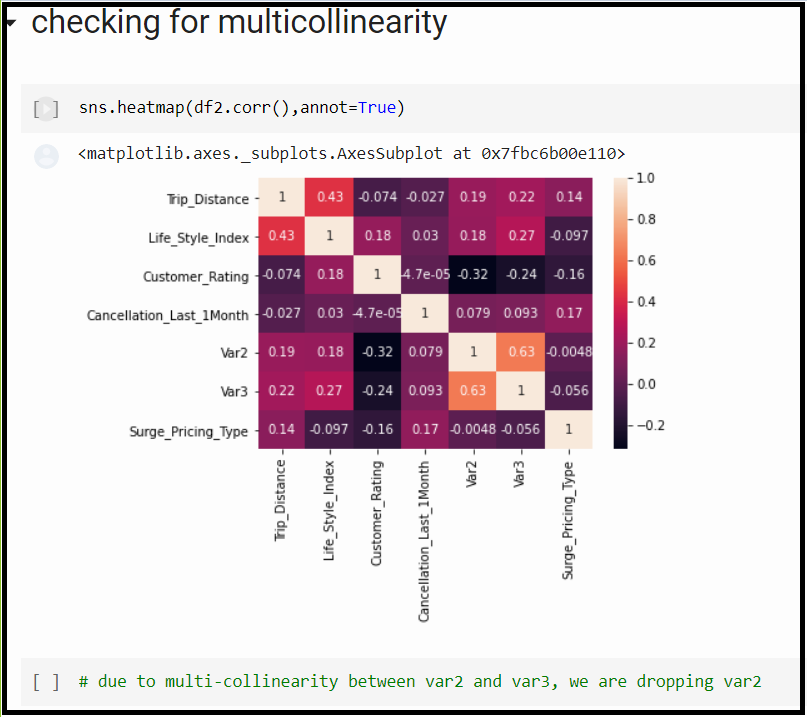


**INFERENCE:**

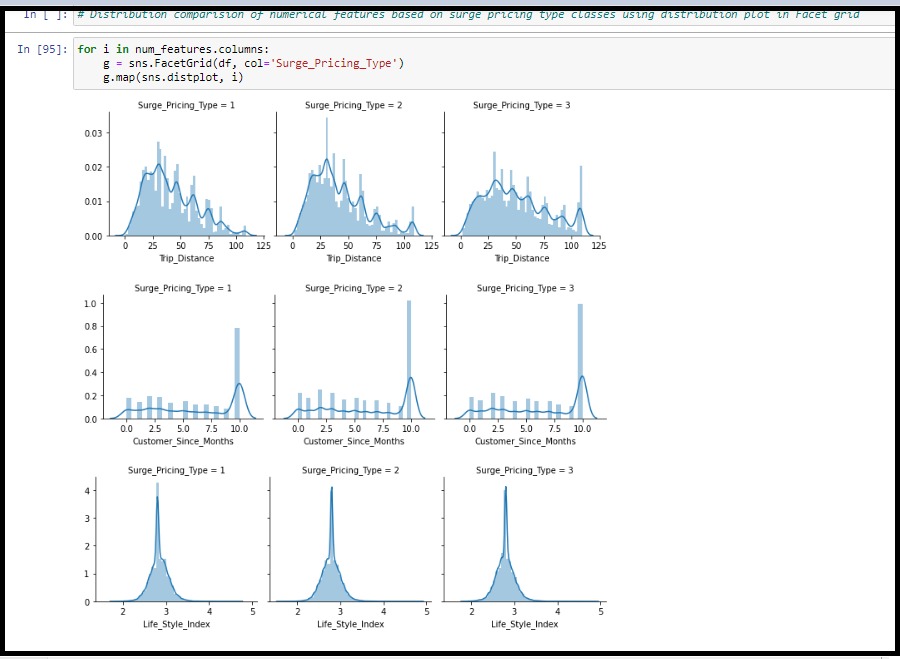
* Yes, we improved significantly on the benchmark.
* With comparison with the base model the f1-weighted score is improved with (6% efficiency) and the bias error is also reduced (6%).

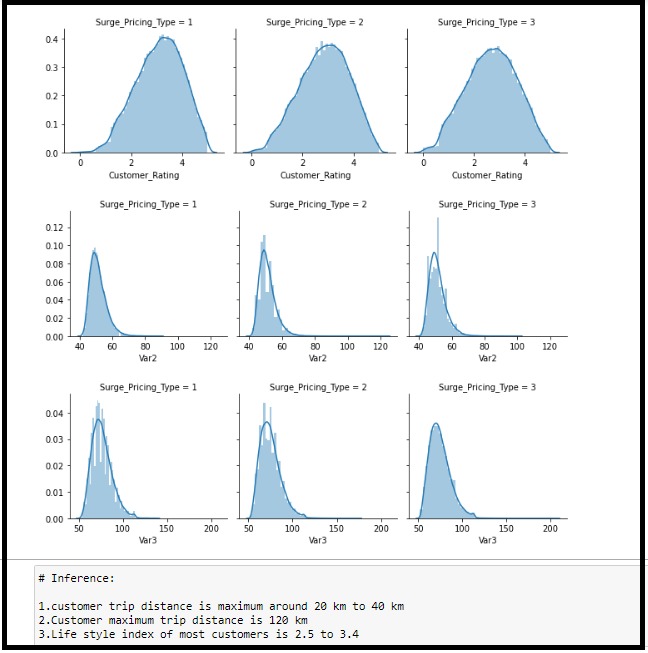
**VISUALIZATION:**

* Multicollinearity
* Univariant Analysis
* Bivariant Analysis
* Multi Variant Analysis
* TABLEAU Analysis.

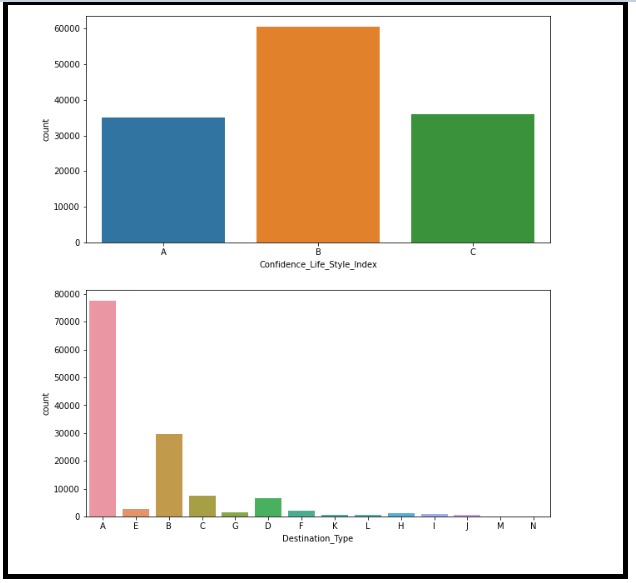


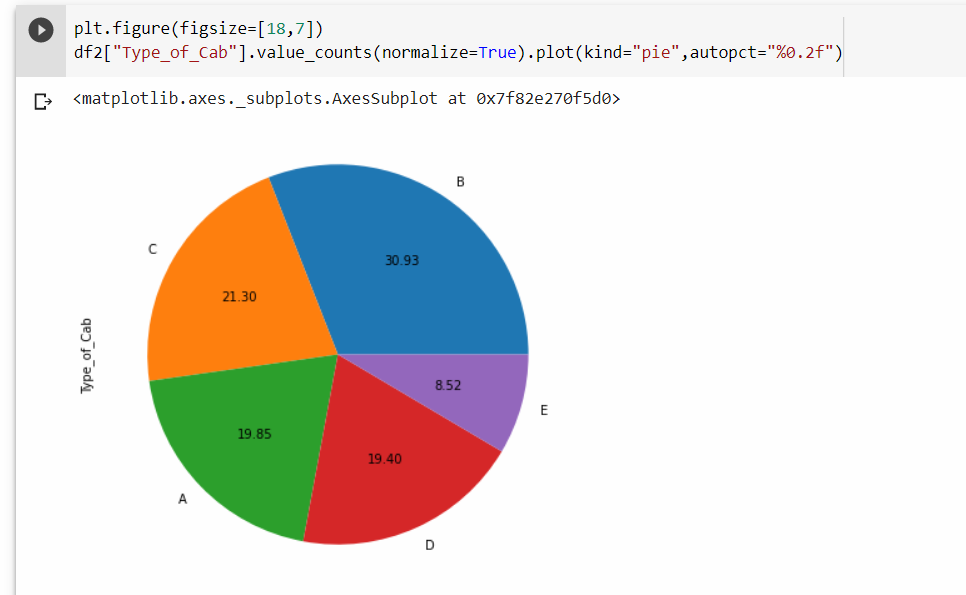
**UNIVARIANT ANALYSIS:NUMERICAL**

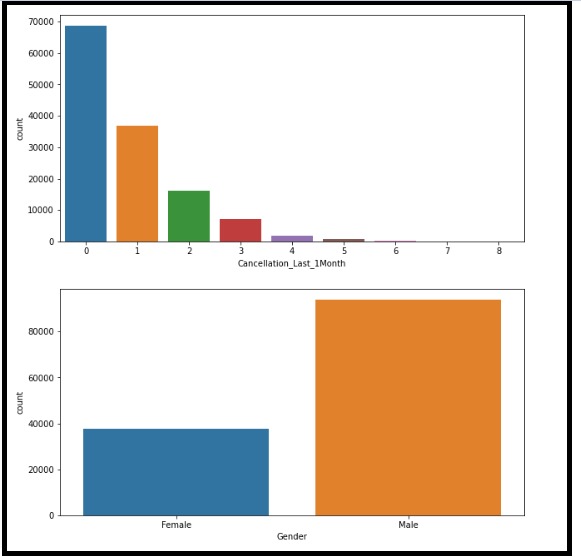
****

****

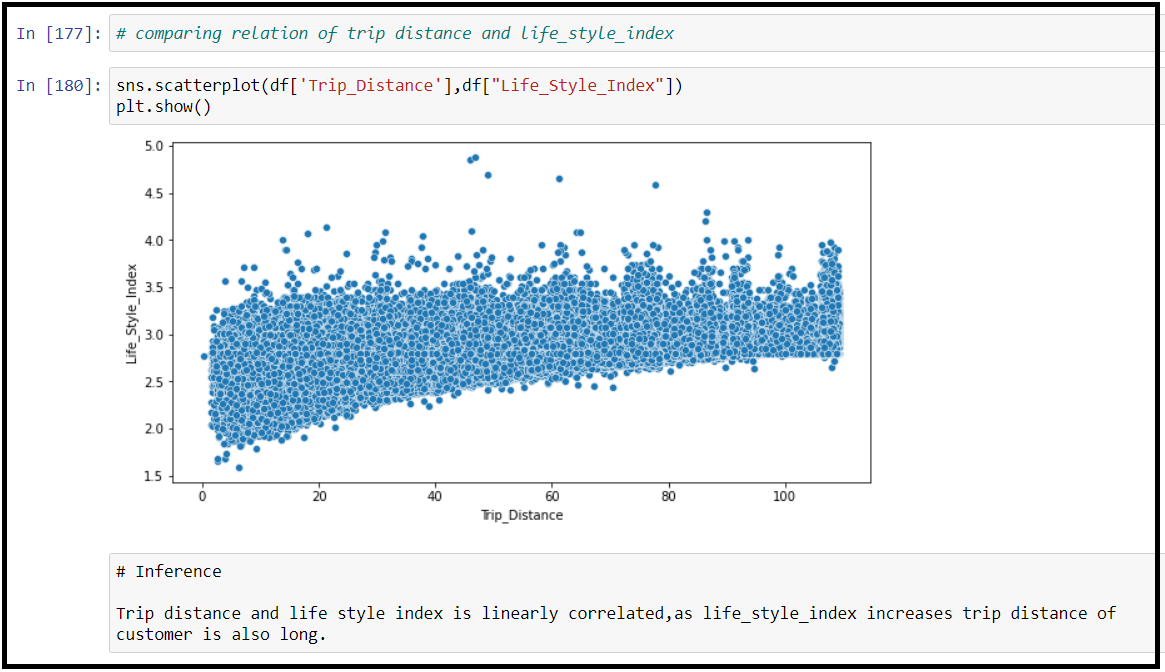
**UNIVARIANT-CATEGORICAL**

****

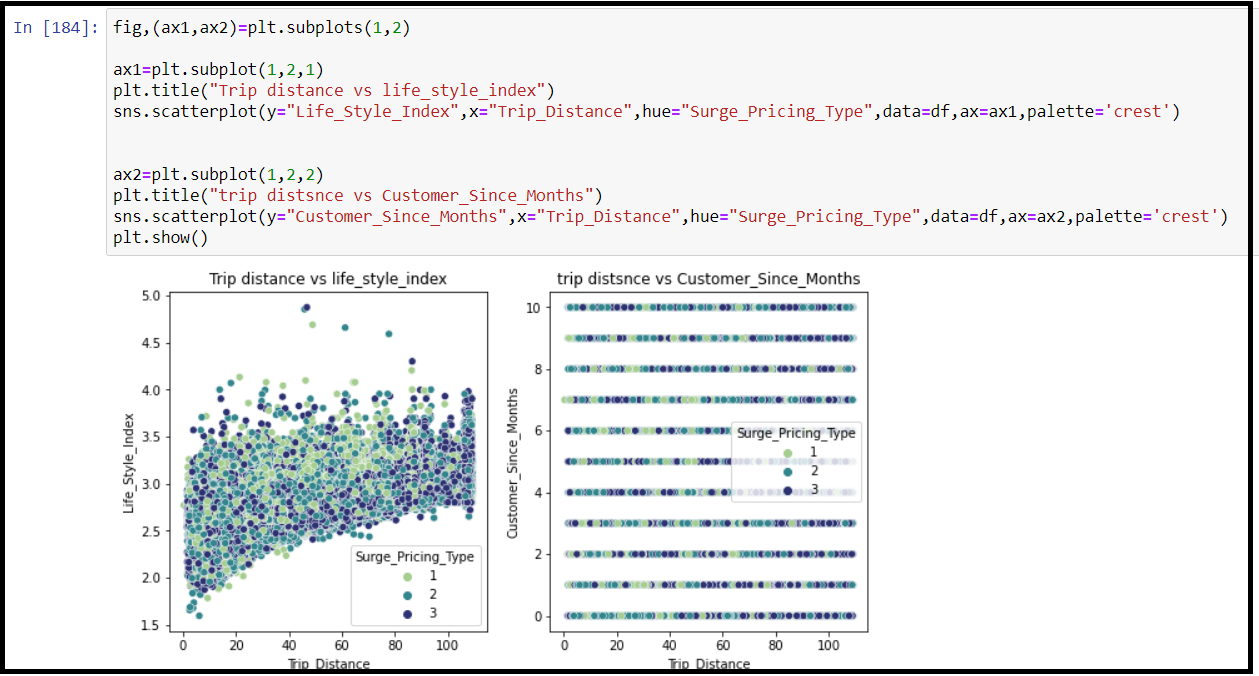


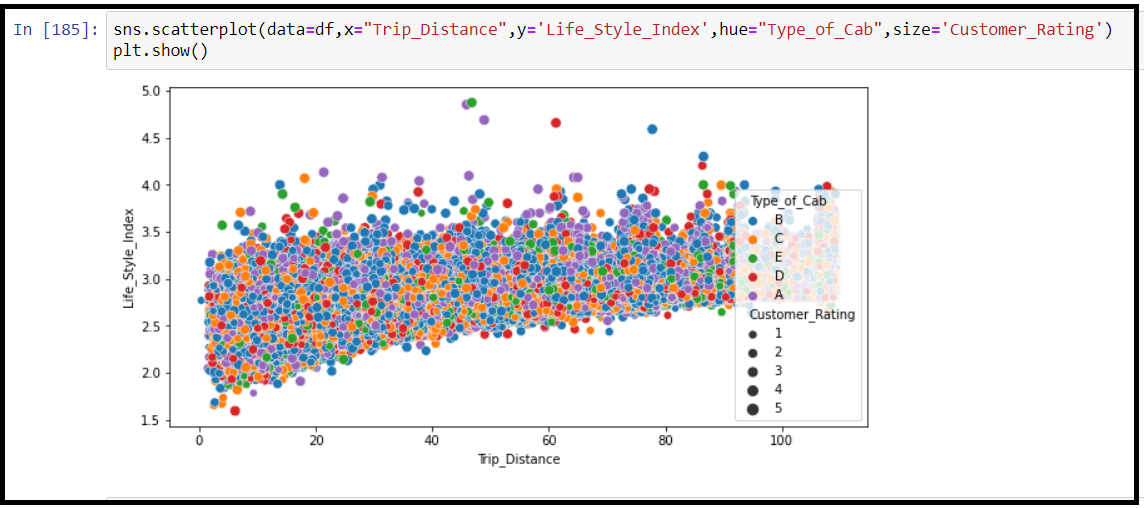
****

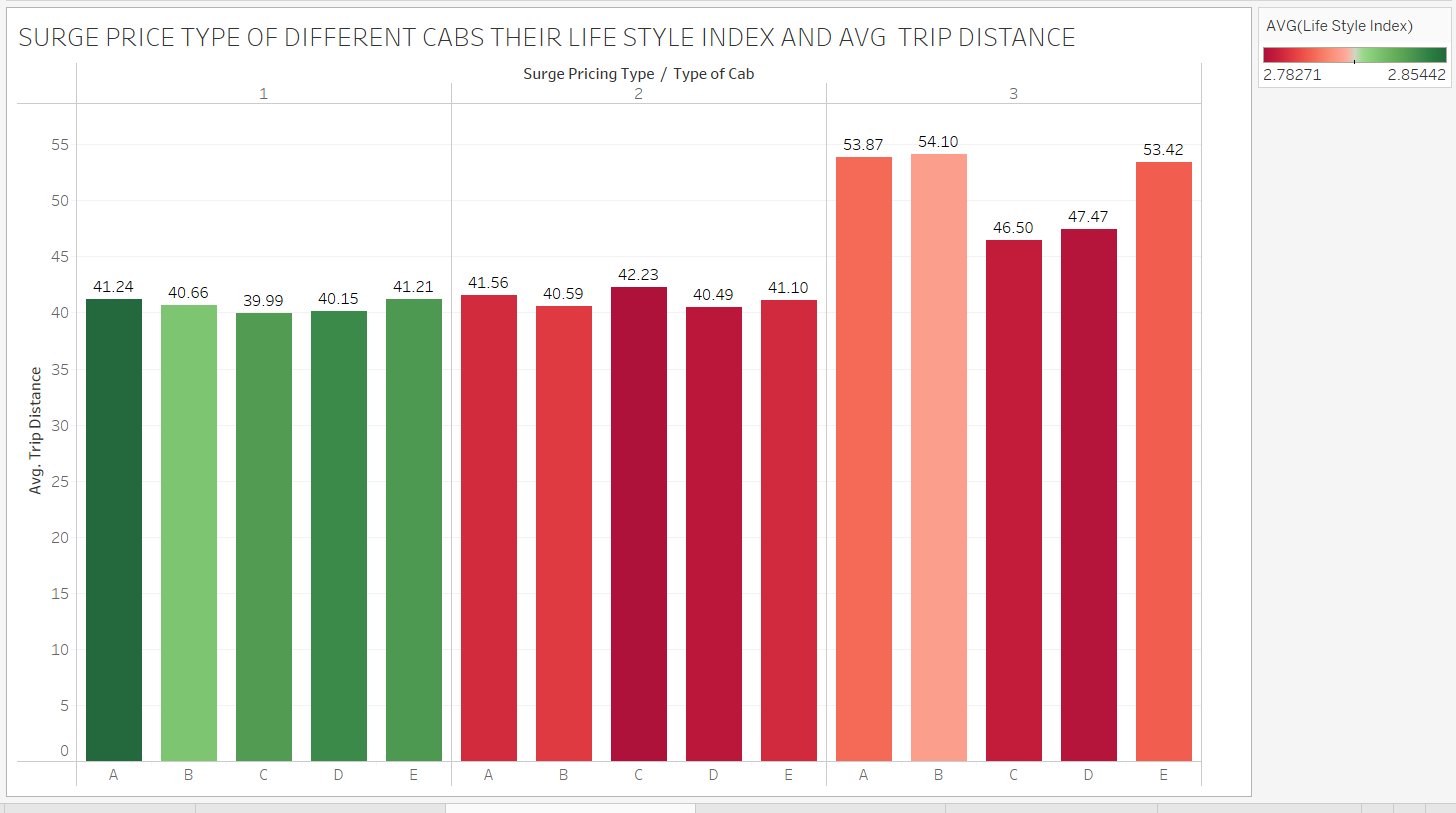
**BIVARIANT ANALYSIS:**

****

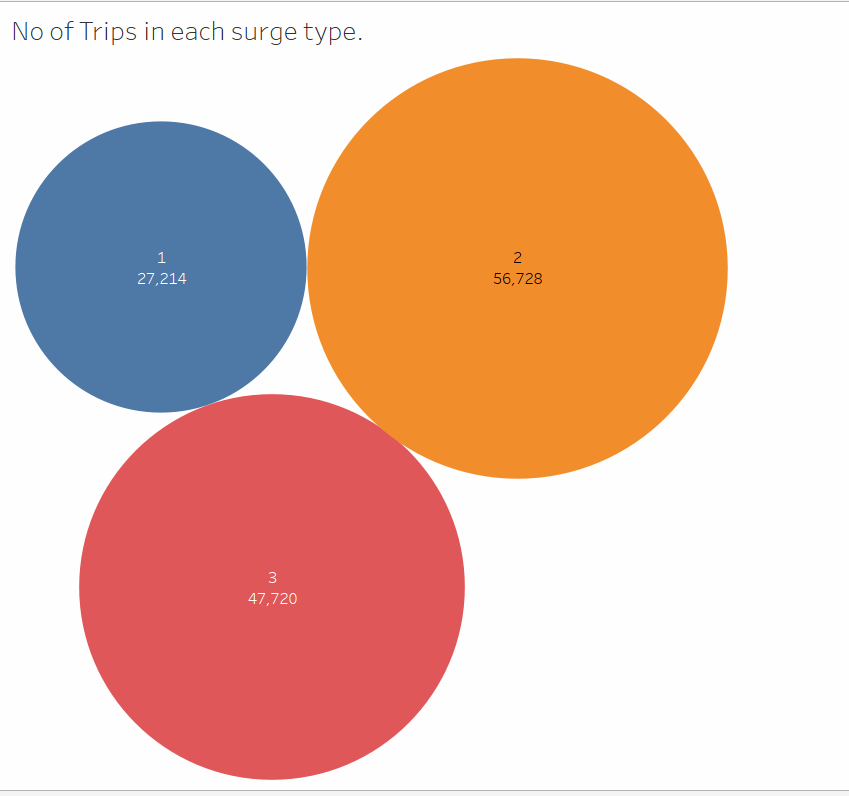
**MULTIVARIANT ANALYSIS:**

****

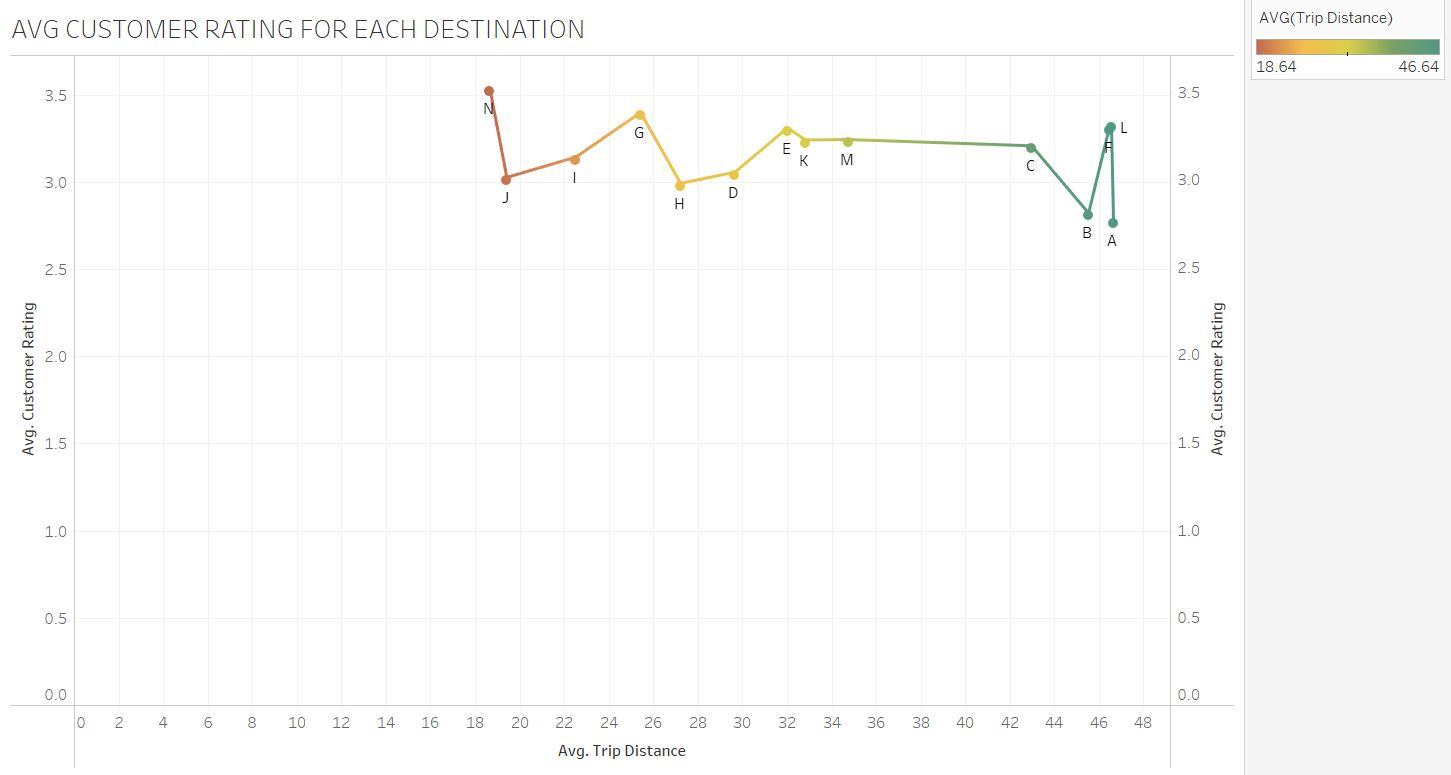
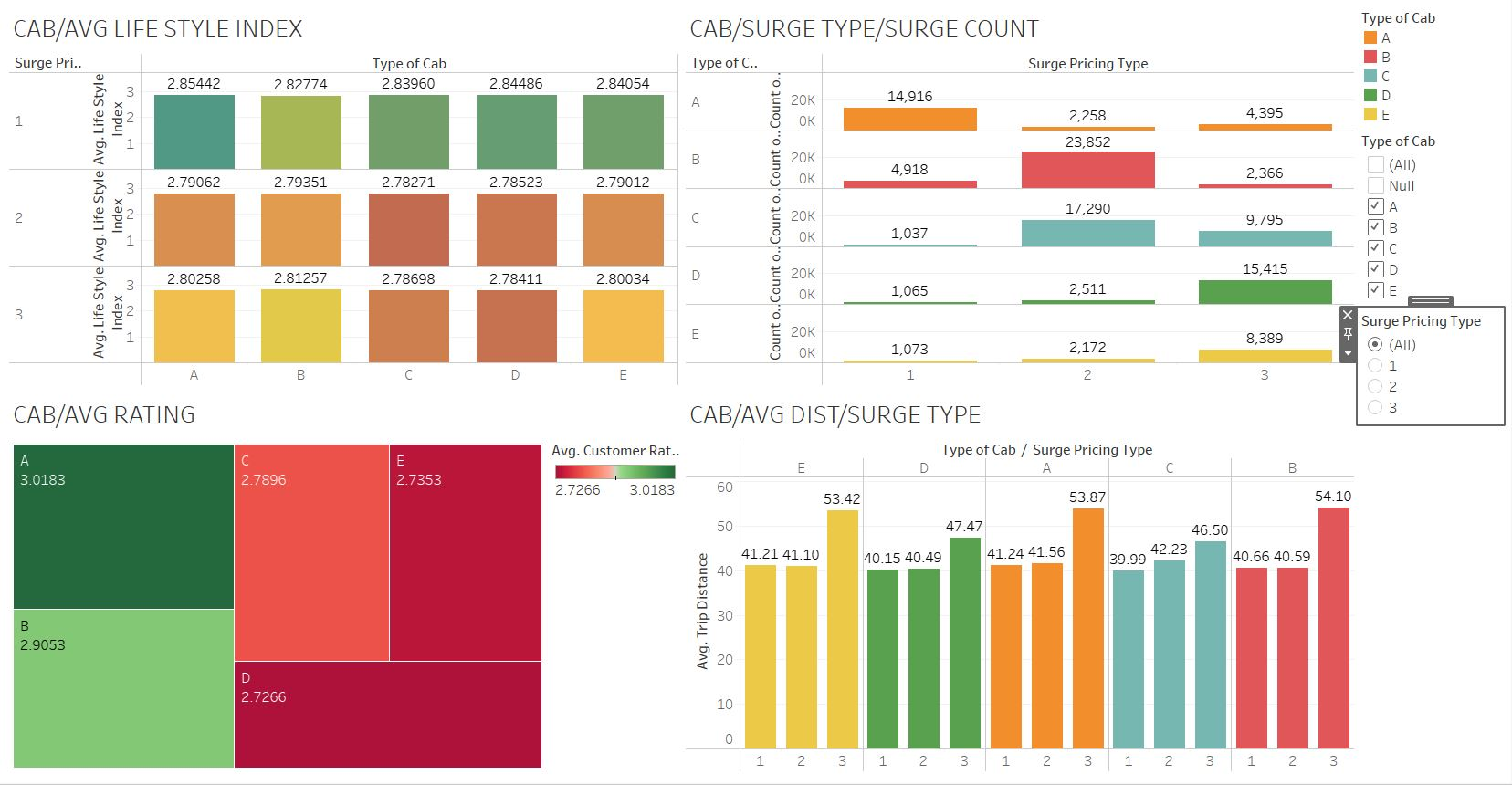
****

**TABLEAU ANALYSIS:**

* Surge\_type1 is having less trip\_distance but average life style index of the customer is high so its assumed that they are capable to pay high surge price even though the rate is high and their aim to reach destination on time.
* Surge\_type\_2 is having almost same trip distance but their life\_style\_index is comparatively lower so they avoid high surge price.
* Even though Surge\_type\_3 is having higher distance since their life\_style index is medium they are capable of paying slightly higher prices.

****

* Most number of people are travelling n surge\_price 2 followed by surge\_price3 then by surge\_price1.
* So we are concluding based on previous analysis surge\_type\_1 price is high,surge\_type\_2 price is low and surge\_type\_3 price is medium.

****

**BUSINESS INSIGHTS:**

Our project is for commercial purpose. We are helping Sigma Cabs, a startup cab aggregator to predict the surge price type with the data they have provided

**IMPLICATIONS:**

* Our f1-Weighted score is only 80% so while in production the model can emit some false predictions which can affect the business.
* In order to get more accurate predictions, we need more data for the model to train.

**LIMITATIONS:**

* Var1, Var2, Var3 is masked variables so we don’t know exactly what role it plays in the data.
* Target Variable is also masked due to confidentiality issues.
* There is considerable multicollinearity between var2 and var3 which can affect the performance of the model in production.
* With general data cleaning the model performance was poor.

**CLOSING REFLECTIONS:**

* We got exposure real world data and its implications.
* We learnt about the process of cab aggregation.
* We learnt about the Simple imputer library in sklearn.
* After doing in depth analysis in Tableau, we learned more about the target variable which is surge pricing type.

**MODEL DEPLOYMENT:**

Deployment is done using Flask, HTML, CSS.

We have attached the deployment file along with this report for your reference.

**FUTURE SCOPE:**

* We will ask to the client for more information regarding the price, Traffic level at the time of the trip.
* We will use more advanced modelling techniques in order to make the prediction more accurate.