By
Jash Bhatia-60009200073

ML-1 Mini Project

Instructor: - Dr. Kriti Srivastava

3rd June 2022

TOPIC

Effective Feedback Management

INTRODUCTION

On the topic of Effective Feedback Management, we choose a dataset on the online delivery system with feedback given on its services by the residents of Bengaluru. This dataset, collected from the residents of Bengaluru which studies on factors which are contributing to the demand of food delivery in the city. The goal is to see if we can predict customer churn. We will perform some visualizations before we go on to pre-process the data, after which we will implement classification models.

DATA DESCRIPTION

The given dataset contains some personal details about the consumer like the consumer's age, gender, marital status, occupation, monthly income, educational qualifications, family size and pin code. The dataset also gives you the consumer's preferred medium of delivery along with what meal of the day they prefer ordering the most. The dataset also mentions if the consumer finds it easy and convenient and time saving, whether or not they would like more restaurant choices, easy payment options, more offers and discounts. It also asks the consumers about the quality of the food, tracking system. The consumers also give their take on if they prefer self cooking to ordering online, are they concerned about the health concerns associated with eating outside food daily, poor hygiene of the food delivered from the restaurants, if they have had a previous bad experience or not with regard to unavailability, unaffordability, long delivery time, delay of delivery person getting assigned, delay of the delivery person picking up food, wrong food delivered, if there was a missing item issue, if the order was placed by mistake. They were also asked the order time of this mishap, the maximum waiting time they have had to incur. If they had

a residence with a busy location, if the google maps correctly located their residence, if the condition of roads leading to the residence was good. How important was delivery order time for them, the quality of package delivered, the politeness of the delivery man, freshness of the food, temperature of the food, good taste and good quantity of the food. Finally they gave their final reviews on what they felt could be better.

You can see this is a detailed dataset with over 388 reviews by consumers that will help in predicting the customer churn for the delivery app.

An example of the dataset is given below

```
[ ] data.head()
```

Age	Gender	Marital Status	Occupation	Monthly Income	Educational Qualifications	Family size	latitude	longitude	Pin code	...	Less Delivery time	High Quality of package	Number of calls	Politeness	Freshness	Temperature	Good Taste	Good Quantity	Output	Reviews
20	Female	Single	Student	No Income	Post Graduate	4	12.9766	77.5993	560001	...	Moderately Important	Moderately Important	Moderately Important	Moderately Important	Moderately Important	Moderately Important	Moderately Important	Moderately Important	Yes	Nil\n
24	Female	Single	Student	Below Rs.10000	Graduate	3	12.9770	77.5773	560009	...	Very Important	Very Important	Very Important	Very Important	Very Important	Very Important	Very Important	Very Important	Yes	Nil
22	Male	Single	Student	Below Rs.10000	Post Graduate	3	12.9551	77.6593	560017	...	Important	Very Important	Moderately Important	Very Important	Very Important	Important	Very Important	Moderately Important	Yes	Many a times payment gateways are an issue, so...
22	Female	Single	Student	No Income	Graduate	6	12.9473	77.5616	560019	...	Very Important	Important	Moderately Important	Very Important	Very Important	Very Important	Very Important	Important	Yes	nil
22	Male	Single	Student	Below Rs.10000	Post Graduate	4	12.9850	77.5533	560010	...	Important	Important	Moderately Important	Important	Important	Important	Very Important	Very Important	Yes	NIL

ws × 55 columns

```
[ ] data.describe()
```

	Age	Family size	latitude	longitude	Pin code
count	388.000000	388.000000	388.000000	388.000000	388.000000
mean	24.628866	3.280928	12.972058	77.600160	560040.113402
std	2.975593	1.351025	0.044489	0.051354	31.399609
min	18.000000	1.000000	12.865200	77.484200	560001.000000
25%	23.000000	2.000000	12.936900	77.565275	560010.750000
50%	24.000000	3.000000	12.977000	77.592100	560033.500000
75%	26.000000	4.000000	12.997025	77.630900	560068.000000
max	33.000000	6.000000	13.102000	77.758200	560109.000000

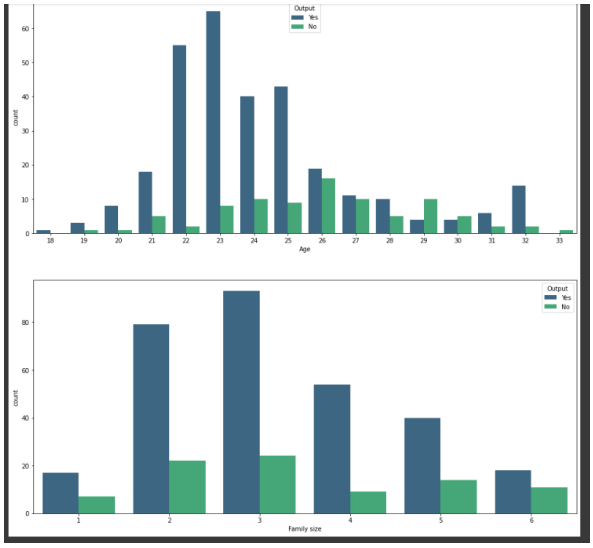
```
data.info()
```

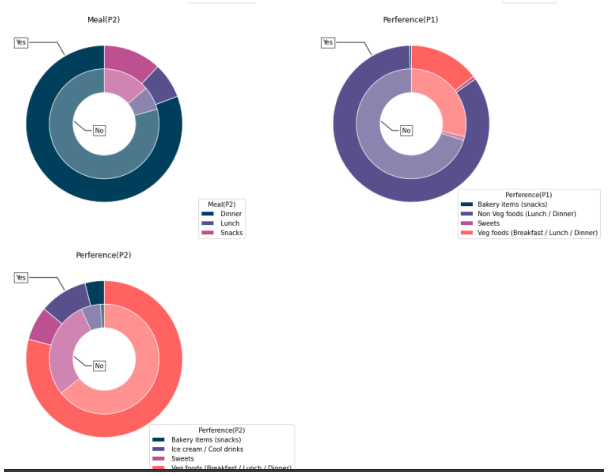
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 388 entries, 0 to 387
Data columns (total 55 columns):
 #   Column                                Non-Null Count  Dtype  
---  ------                                --------------  -----  
 0   Age                                   388 non-null    int64  
 1   Gender                                388 non-null    object 
 2   Marital Status                        388 non-null    object 
 3   Occupation                            388 non-null    object 
 4   Monthly Income                        388 non-null    object 
 5   Educational Qualifications            388 non-null    object 
 6   Family size                           388 non-null    int64  
 7   latitude                              388 non-null    float64
 8   longitude                             388 non-null    float64
 9   Pin code                              388 non-null    int64  
 10  Medium (P1)                           388 non-null    object 
 11  Medium (P2)                           388 non-null    object 
 12  Meal(P1)                              388 non-null    object 
 13  Meal(P2)                              388 non-null    object 
 14  Perference(P1)                        388 non-null    object 
 15  Perference(P2)                        388 non-null    object 
 16  Ease and convenient                   388 non-null    object 
 17  Time saving                           388 non-null    object 
 18  More restaurant choices               388 non-null    object 
 19  Easy Payment option                   388 non-null    object 
 20  More Offers and Discount              388 non-null    object 
 21  Good Food quality                     388 non-null    object 
 22  Good Tracking system                  388 non-null    object 
 23  Self Cooking                          388 non-null    object 
 24  Health Concern                        388 non-null    object 
 25  Late Delivery                         388 non-null    object 
 26  Poor Hygiene                          388 non-null    object 
 27  Bad past experience                   388 non-null    object 
 28  Unavailability                        388 non-null    object 
 29  Unaffordable                          388 non-null    object 
 30  Long delivery time                    388 non-null    object 
 31  Delay of delivery person getting assigned 388 non-null    object 
 32  Delay of delivery person picking up food  388 non-null    object 
 33  Wrong order delivered                 388 non-null    object 
 34  Missing item                          388 non-null    object 
 35  Order placed by mistake               388 non-null    object 
 36  Influence of time                     388 non-null    object 
 37  Order Time                            388 non-null    object 
 38  Maximum wait time                     388 non-null    object 
 39  Residence in busy location            388 non-null    object 
 40  Google Maps Accuracy                  388 non-null    object 
 41  Good Road Condition                   388 non-null    object 
 42  Low quantity low time                 388 non-null    object 
 43  Delivery person ability               388 non-null    object 
 44  Influence of rating                   388 non-null    object 
 45  Less Delivery time                    388 non-null    object 
 46  High Quality of package               388 non-null    object 
 47  Number of calls                       388 non-null    object 
 48  Politeness                            388 non-null    object 
 49  Freshness                             388 non-null    object 
 50  Temperature                           388 non-null    object 
 51  Good Taste                            388 non-null    object 
 52  Good Quantity                         388 non-null    object 
 53  Output                                388 non-null    object 
 54  Reviews                               388 non-null    object 
dtypes: float64(2), int64(3), object(50)
memory usage: 166.8+ KB
```

DATA ANALYSIS

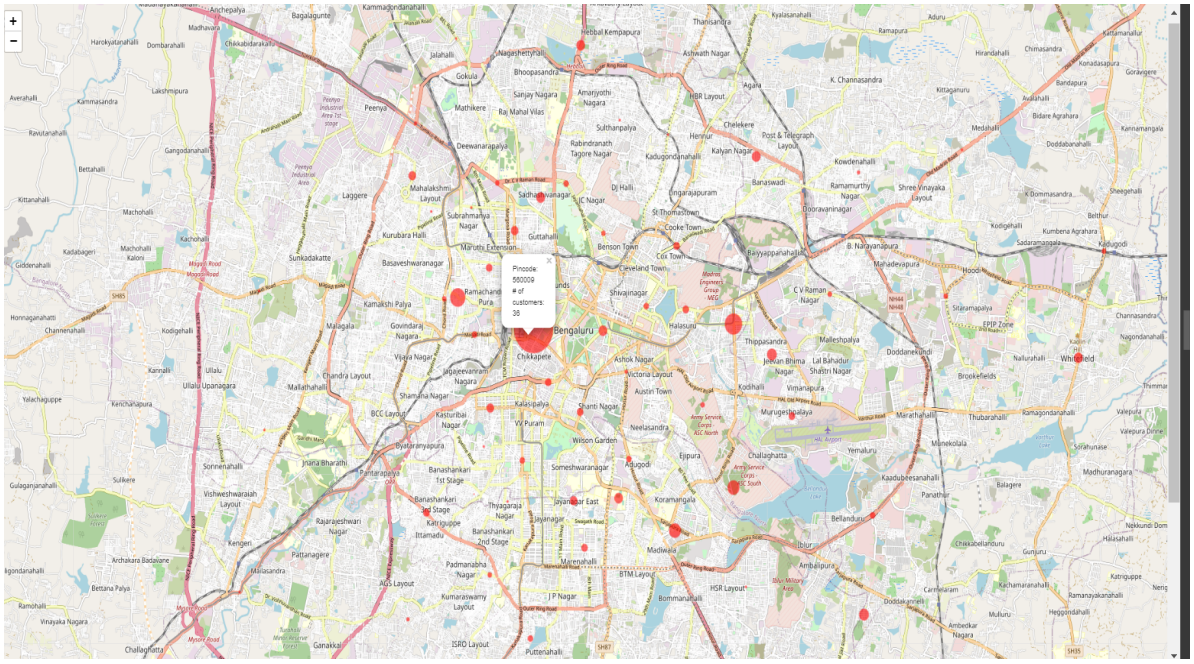
Since the data is 400*55 we have performed data visualizations and and pre processed the data

We have analyzed the data using such graphs and then dropped the data that we didnt find useful

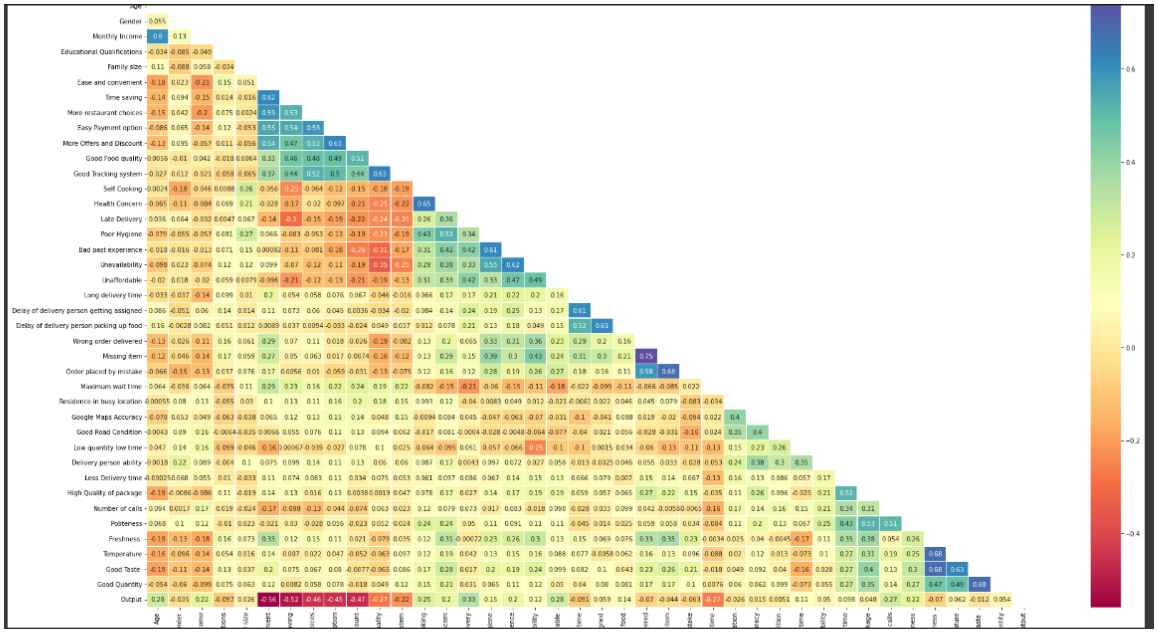




We used geospatial analysis to use the latitude longitude data that has been provided in the dataset



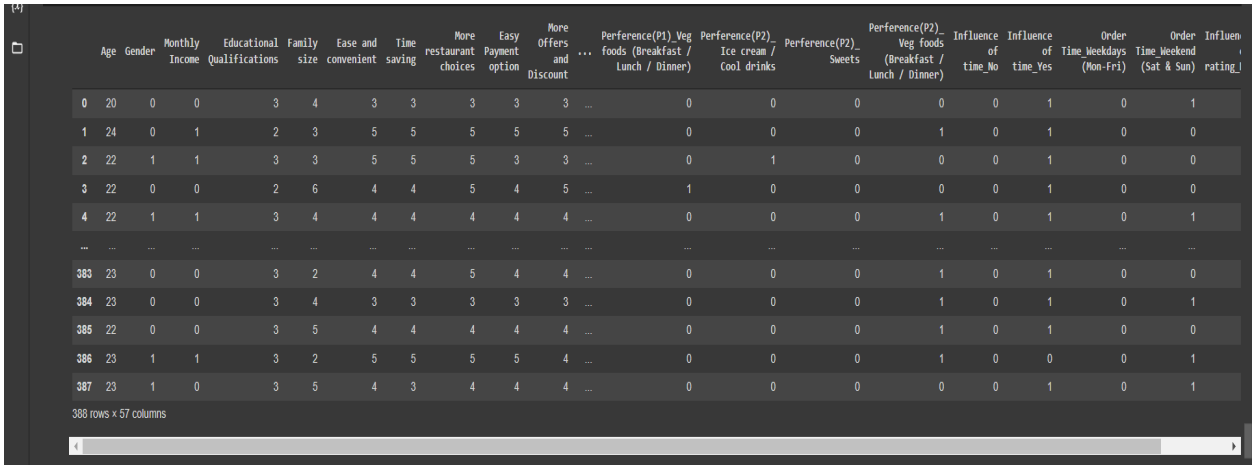
We showed using this map how many orders have been delivered in the same locality. Since this particular data doesn't help in predicting the customer churn we drop the data.



We notice that not only are features from 'Ease and convenient' to 'Good Tracking system' correlated with our output variable, but also with each other. In fact, if we observe the diagonals, we can see which features are correlated with each other, which gives us an insight as to how we can compress the multiple dimensions in our variables. One of the popular dimensionality reduction methods, principal component analysis will not help in this case because the variables

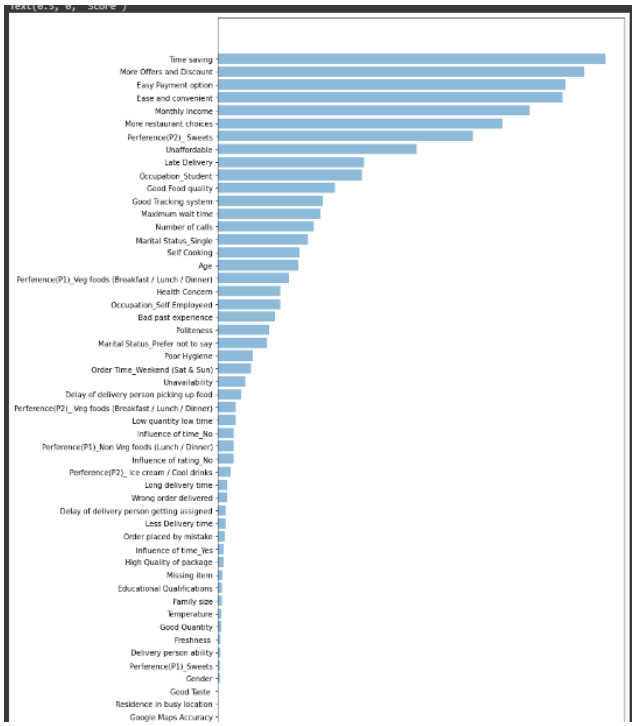
are of ordinal categorical nature rather than continuous. We will use other feature selection methods later in the data pre-processing stage

Some of the pre-processing has been done in the visualization step itself, where we converted likert features to an ordinal scale. There are a few features remaining which are still categorical. We will convert these to dummy variables now



	Age	Gender	Monthly Income	Educational Qualifications	Family size	Ease and convenient	Time saving	More restaurant choices	Easy Payment option	More Offers and Discount	...	Perference(P1)_Veg foods (Breakfast / Lunch / Dinner)	Perference(P2)_ Ice cream / Cool drinks	Perference(P2)_ Sweets	Perference(P2)_ Veg foods (Breakfast / Lunch / Dinner)	Influence of time_No	Influence of time_Yes	Order Time_Weekdays (Mon-Fri)	Order Time_Weekend (Sat & Sun)	Influen rating_
0	20	0	0	3	4	3	3	3	3	3	...	0	0	0	0	0	1	0	1	
1	24	0	1	2	3	5	5	5	5	5	...	0	0	0	1	0	1	0	0	
2	22	1	1	3	3	5	5	5	3	3	...	0	1	0	0	0	1	0	0	
3	22	0	0	2	6	4	4	5	4	5	...	1	0	0	0	0	1	0	0	
4	22	1	1	3	4	4	4	4	4	4	...	0	0	0	1	0	1	0	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
383	23	0	0	3	2	4	4	5	4	4	...	0	0	0	1	0	1	0	0	
384	23	0	0	3	4	3	3	3	3	3	...	0	0	0	1	0	1	0	1	
385	22	0	0	3	5	4	4	4	4	4	...	0	0	0	1	0	1	0	0	
386	23	1	1	3	2	5	5	5	5	4	...	0	0	0	1	0	0	0	1	
387	23	1	0	3	5	4	3	4	4	4	...	0	0	0	0	0	1	0	1	

388 rows x 57 columns



As a starting point, we will select the top 20 features with the highest feature scores and see if we need to drop any variables later due to overfitting

Reasons to use ML model

We have used three models on the given dataset:-

1. Logistic Regression
2. Random Forest
3. K-Nearest Neighbors

Since the target variable is categorical, we have used classification models on the given dataset.

ALGORITHM

In the three algorithms mentioned above we have preferred Random forest algorithm and KNN clustering algorithm as it gives a higher accuracy than logistic regression algorithm.

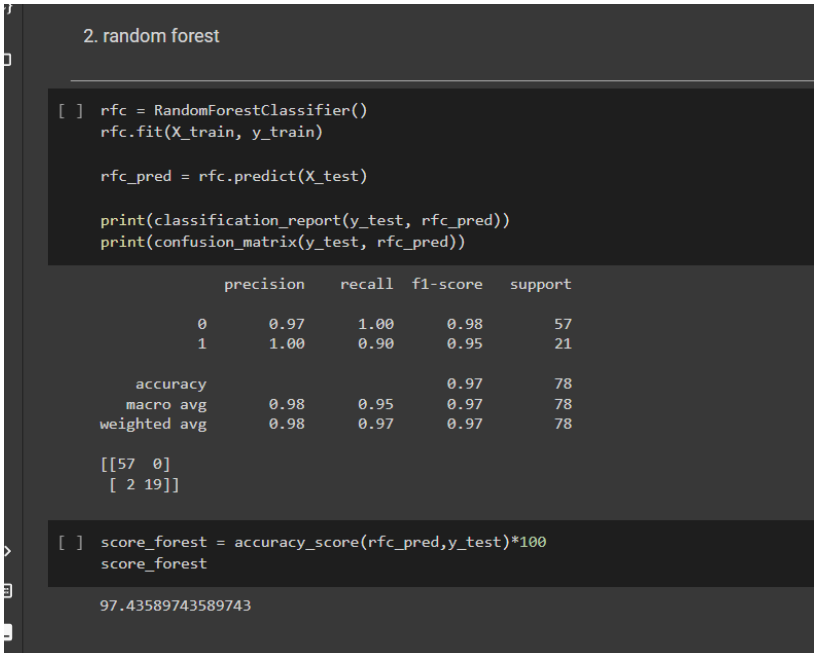
Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Random forest algorithm avoids and prevents overfitting by using multiple trees. The results are not accurate. This gives accurate and precise results. Decision trees require low computation, thus reducing time to implement and carrying low accuracy.

Therefore, We prefer the Random forest algorithm on the given dataset.

Result Analysis

We achieved a accuracy of 97.43% to predict customer churn using random forest algorithm



2. random forest

```
[ ] rfc = RandomForestClassifier()
    rfc.fit(X_train, y_train)

    rfc_pred = rfc.predict(X_test)

    print(classification_report(y_test, rfc_pred))
    print(confusion_matrix(y_test, rfc_pred))
```

```
                precision    recall  f1-score   support

             0       0.97      1.00      0.98        57
             1       1.00      0.90      0.95        21

      accuracy                           0.97        78
     macro avg       0.98      0.95      0.97        78
  weighted avg       0.98      0.97      0.97        78

[[57  0]
 [ 2 19]]
```

```
[ ] score_forest = accuracy_score(rfc_pred,y_test)*100
    score_forest
```

```
97.43589743589743
```

Conclusions and Future scope

Out of all models created, Random Forest and K Nearest Neighbors give the best performance. K Nearest Neighbors has the added advantage of being relatively simpler to interpret, without any sacrifice in accuracy.

Another consideration is about the data itself. We see that most of the variables that have a material impact on customer churn relate to broad aspects of ease and convenience. In fact, the subset of features selected in the model could inform business decisions, regarding which aspects of the service could be focused on for minimizing churn, and what user segments should the marketing dollars be spent on.

NOTE:- We ignored the text data, i.e. reviews in this analysis and analyzing with NLP methods could also add to the analysis