CAPSTONE PROJECT

FINAL REPORT

23RD JUNE

Group 6, Jan 2021 Batch

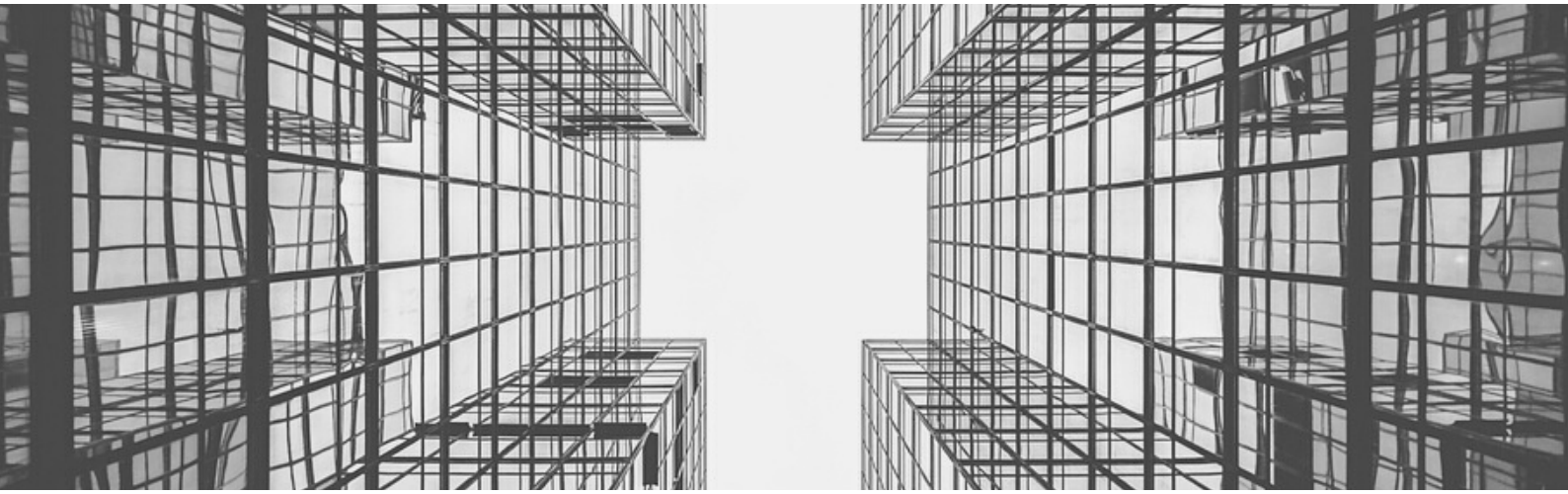
PROJECT SUMMARY

Batch details	PGP - DSE – Jan 2021
Team members	Divya Nandakumar
	Nivedhan S
	Ganesh M
	Ahilapriya SN
Domain of Project	Banking and Finance
Proposed project title	Customer review analysis
Group Number	6
Team Leader	Divya Nandakumar
Mentor Name	Shirude Shashank Pathak

Date: 23rd June 2021

Signature of the mentor **Signature of the team leader**

CONTENTS



OVERVIEW

The Financial sector plays an integral role in driving the economy of a nation. Economic growth and sector success are interlinked and digging deep into the intuitive nuances of the financial data helps in making better-informed decisions for the long run.

As we know that in the present time individuals are more into business. So, accepting an objection from a buyer happens consistently. A customer's grumblings present a bank or any organization with a chance to distinguish and correct explicit issues present in their products or administration. Effective addressal of consumer complaints is an essential part of business management. A good complaint-management strategy will result in the best customer relationship outcome with minimal human-resource investment. Our research hopes to find a correlation between complaints, companies & consumers to refine their products and to better accommodate consumer needs. (Chugani, K , & Ramasubbareddy, 2018)

Business Problem Statement

01. Aim

We aim on interpreting the customer grievances data and assisting the banks in identifying the types of errors. In other words, we want to help the financial institutions to ensure that every customer is satisfied at the end of the day. Our model will help financial institutions in understanding the data, investigate the root cause of the issue and improve customer satisfaction.

02. Objectives

· To predict if there will be any monetary loss from the company side and to emphasize on corresponding areas to save cost associated with the complaint.
· To predict if there will be timely responses for the issue, so that respected team can emphasize more on such complaints to resolve it on time.

03. Approach

- Data Balancing.
- EDA and Pre-processing.
- Classification - Random Forest and XGB Classifier.
- Feature selection and Hyper parameter Tuning.

04. Future Scope

To deploy the model and automate the process.

05. Limitations

The model needs to be updated with time and business.

Topic Survey

01 - Problem Understanding

We have 18 attributes in this dataset that define the issues customers have faced while using financial services and products. Dissatisfied customers will often feel reluctant to use the services in the future and thus hampering customer retention. The key to any successful business relays on one important line ‘Customer is the CEO and it is our utmost priority to make them happy. Through our analysis we aim to bridge the gap between the customers and the company by deeply understanding the root cause of the issue. Customer satisfaction analysis in financial services has seen tremendous growth in recent years, owing to the importance of providing high-quality service to customers.

02 - Current solution to the problem

In recent times, the Finance and banking sector have enabled various approaches to combat customer dissatisfaction and timely addressal of issues. Based on our current objectives mentioned above, we aim to reduce the monetary loss for the company and help them understand the issues that lead to it. We aim to provide solutions to the problem by using Data balancing techniques, EDA and Classification models.

03 - Proposed solution to the problem

While numerous factors can aid the growth of a business, customer satisfaction plays an integral role in the company’s overall success. We use supervised learning classification algorithms and fine tune the hyper-parameters to predict accurately.

04 - Reference

Original owner of data	Consumer Financial Protection Bureau
Data set information	Consumer Complaint Database
Citation	Bureau of Consumer Financial Protection. (2019, May 13). *Consumer Complaint Database*.Retrieved Apr 12, 2021 from Kaggle: https://www.kaggle.com/selener/consumer-complaint-database
Link to web page	https://www.kaggle.com/selener/consumer-complaint-database

OBJECTIVE I

Our objective is to predict if there will be any monetary loss from the company side and to emphasize on corresponding areas to save cost associated with the complaint.

Our target column had various sub levels, Keeping our objective in mind and to get some meaningful inferences we decided to join the suitable levels and drop the rest.

We clubbed Closed with non-monetary relief, Closed with explanation, Closed without relief and Closed under the category **'Closed without monetary relief'**.

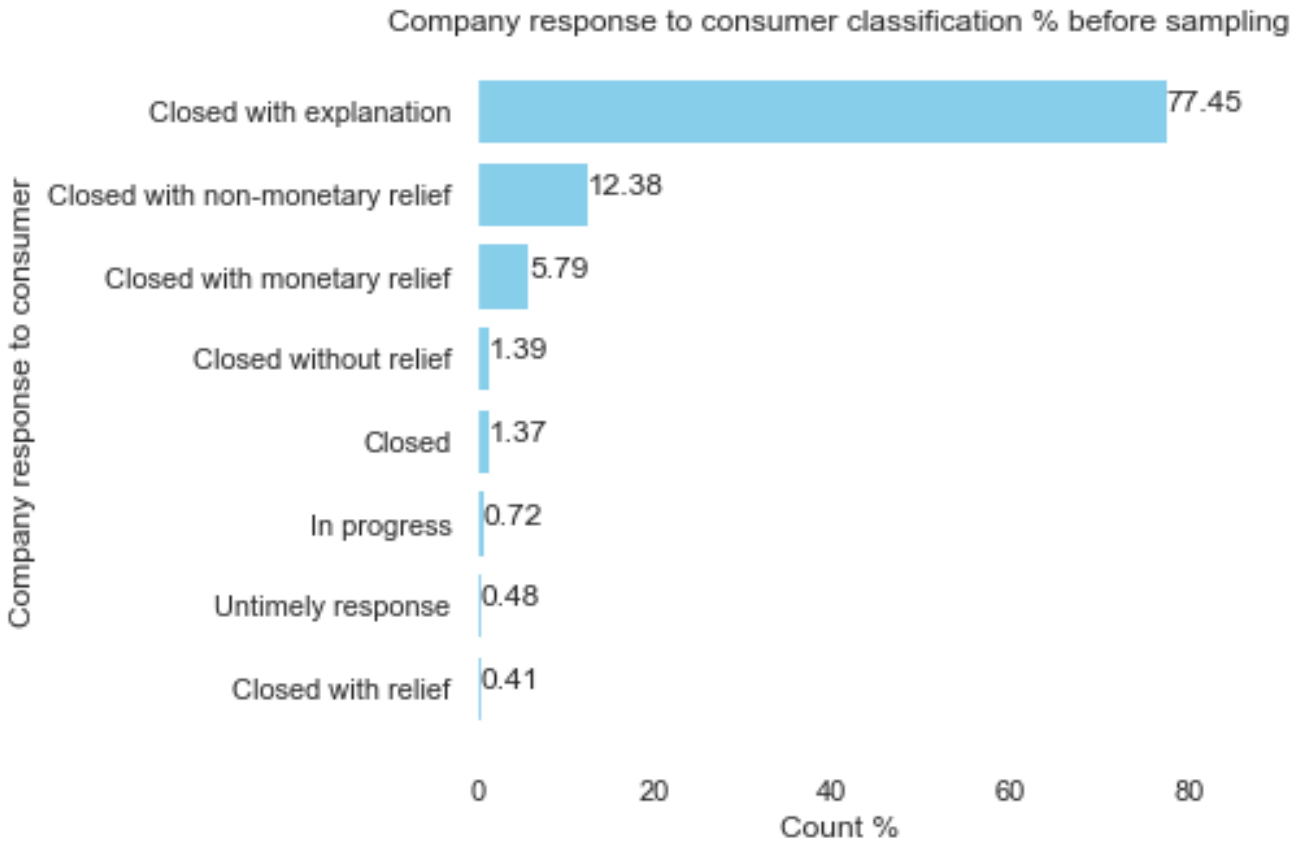


Figure 1: Before applying random sampling technique

We clubbed Closed with monetary relief and closed with relief under the category **'Closed with monetary relief'**.

Our target column now has 2 levels - closed with monetary relief and closed without monetary relief.

After our basic preprocessing of the data, we realised our target column did not have equal representation of the levels.

To overcome this issue, we applied random sampling technique and selected equal number of rows in each column.

The same is representated in the graph below.

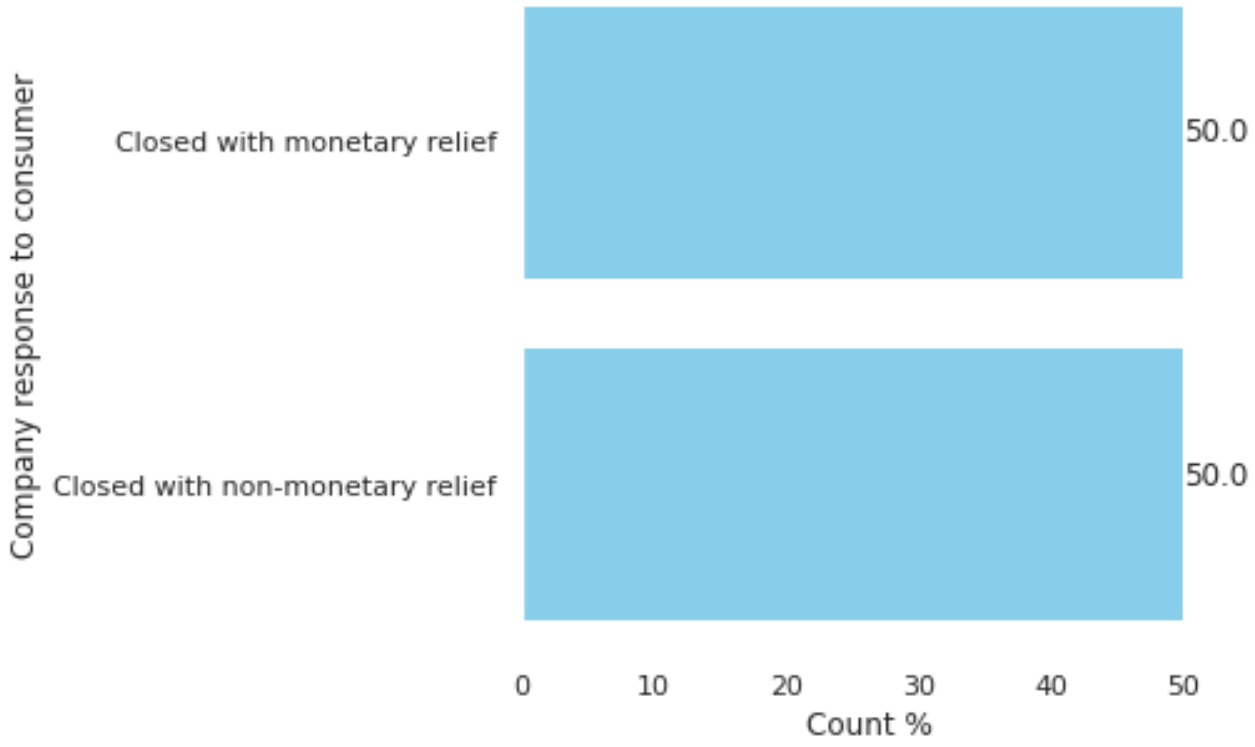


Figure 2: After applying random sampling technique

PREPROCESSING

For the purpose of our objective, we created a data subset using random sampling technique.

After that, our target columns had equal representation of the levels.

The final dataset contains 111338 rows and 14 columns. It was prepared for further analysis by removing or modifying the data that has been duplicated, incorrect, incomplete or irrelevant.

NULL IMPUTATION

Null values have been treated accordingly with respect to the null count in the column;

Column Names	Missing (%)	Handling Method
Sub-product	19.12	Conditional Imputing
Sub-issue	45.79	Column Dropped
State	1.83	Mode Imputation
Consumer consent provided?	37.34	Column Dropped
Consumer disputed?	45.13	Column Dropped

FEATURE ENGINEERING

1. 'State' column had 62 entries which was then clubbed based on the US map into regions having 6 levels.

2. '**Duration'** column had been created by subtracting **'Date sent to company'** and **'Date received**'. This columns basically talks about the time taken for the filed response to reach the company.

3. Post which columns such as 'State', 'Zip Code', 'Date sent to company' and 'Date received' were dropped.

4. The identifier column ('Complaint ID') was dropped from further analysis.

EDA

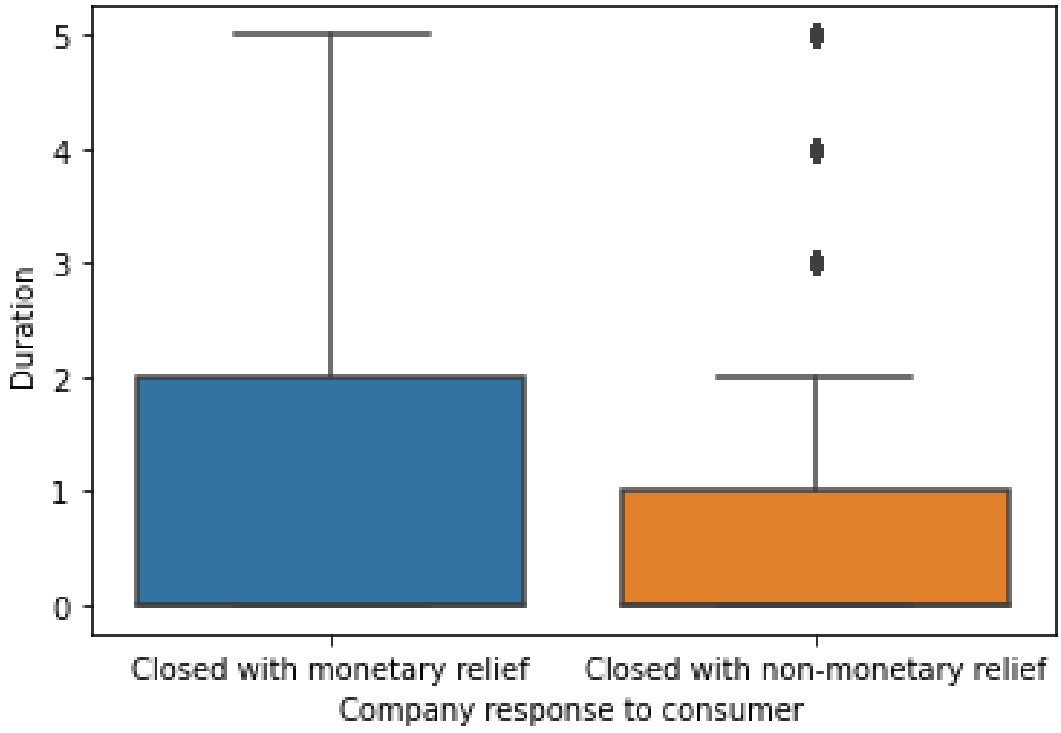


Figure 3: Boxplot of duration with the target column

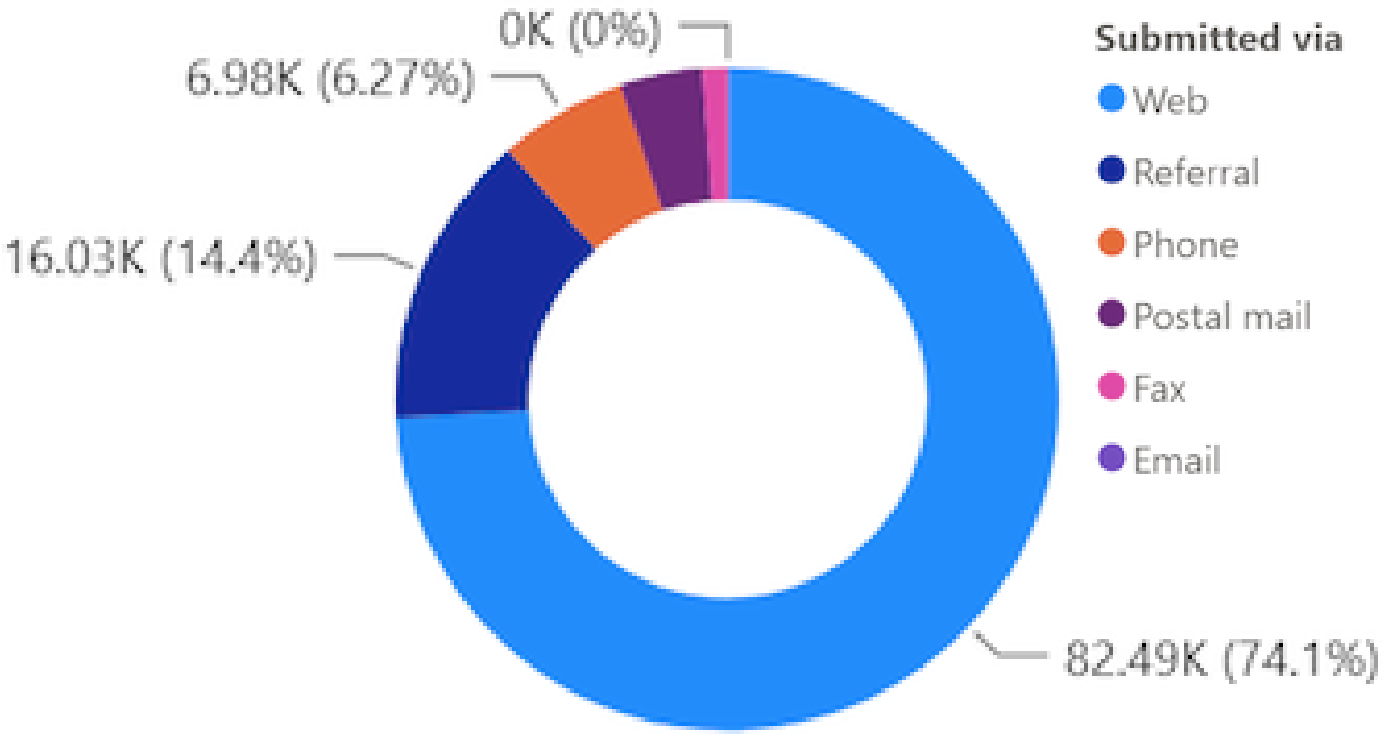


Figure 4: Mode of filing complaints

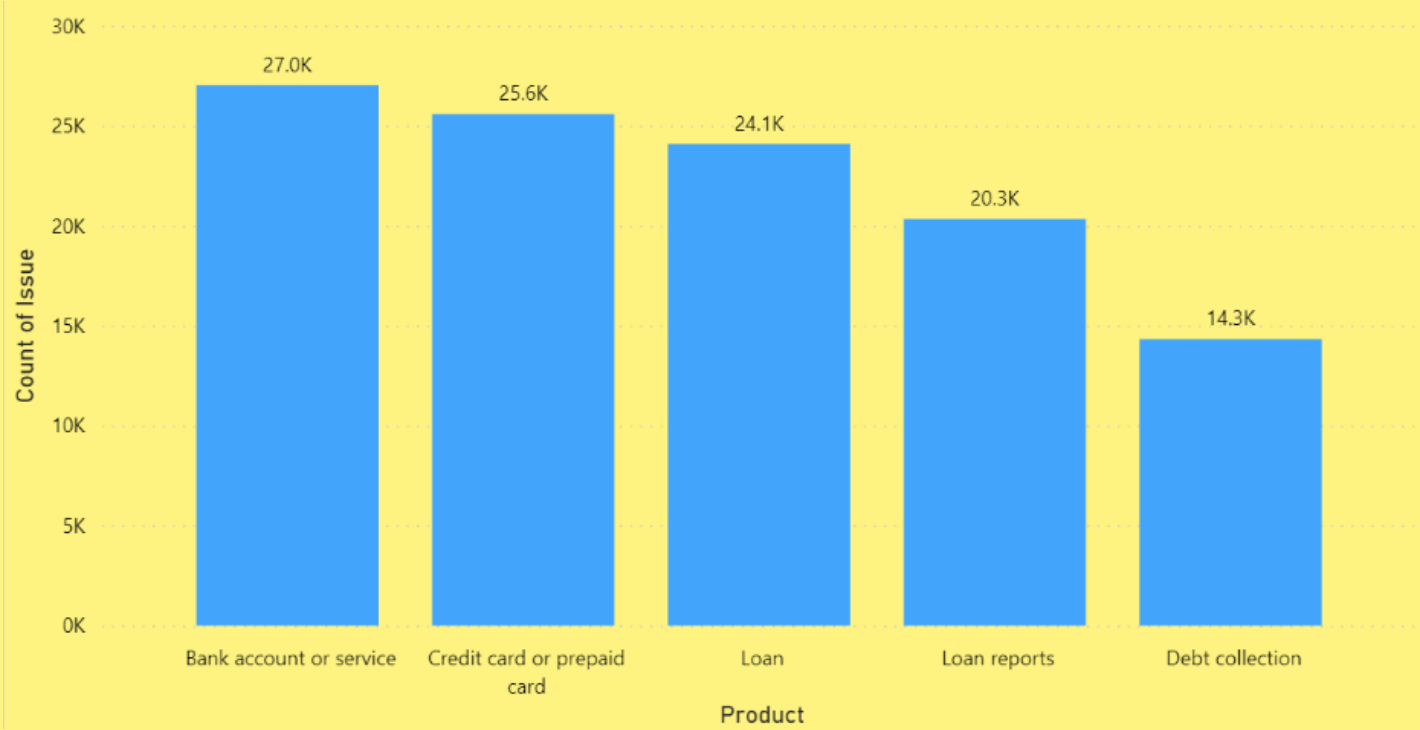


Figure 5: Count of issues by product

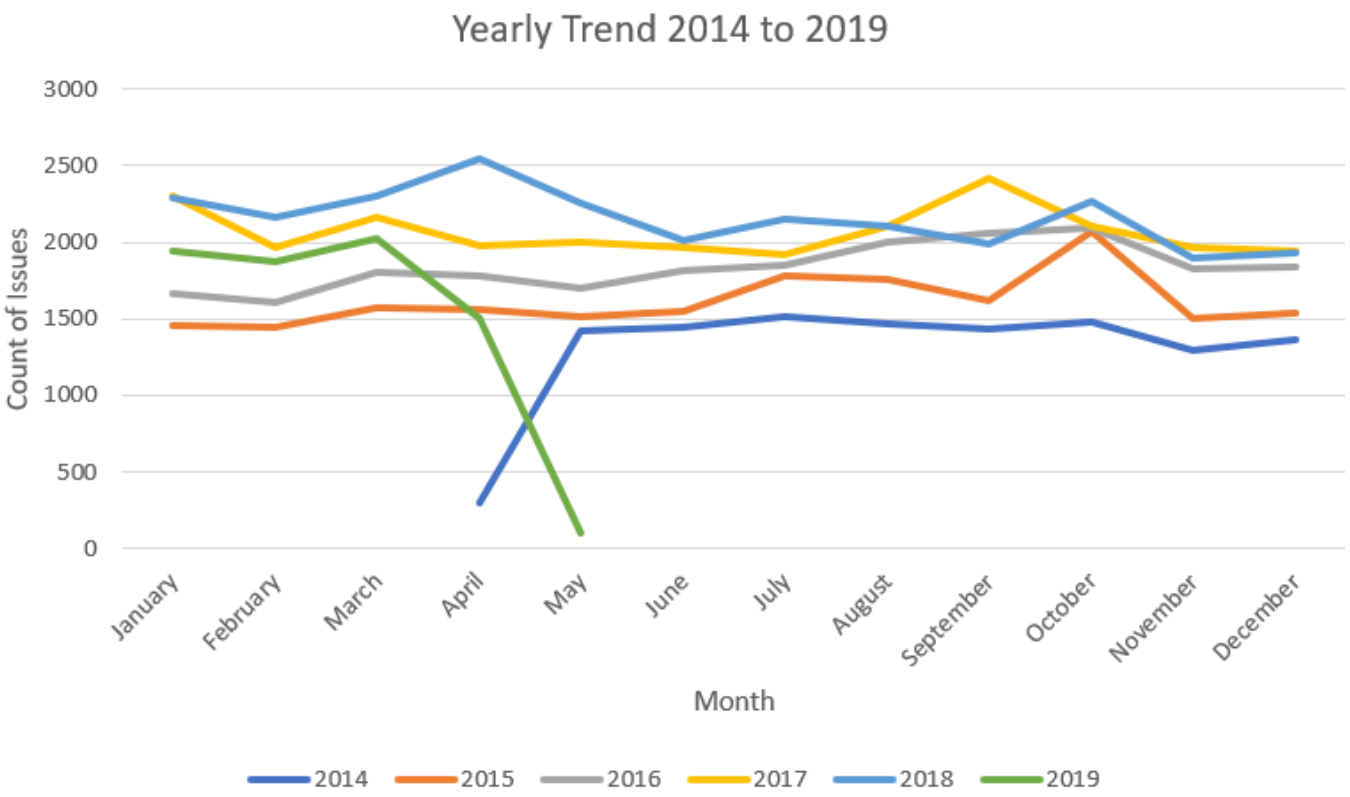


Figure 6: Count of issues yearly trend

ENCODING

Each categorically column was encoded with suitable technique as follows:

Column Names	No. of unique values	Technique
Product	5	Factorize Encoding
Sub-Product	61	Frequency Encoding
Issue	78	Frequency Encoding
Company	2490	Frequency Encoding
Submitted via	6	Factorize Encoding
Timely response	2	Factorize Encoding
Company response to consumer	2	Factorize Encoding
Region	6	Factorize Encoding

MODEL BUILDING

We trained our dataset in various classifiers and evaluated the model based on performance metrics such as:

- Accuracy score
- Recall score
- Precision score
- F1 score

	Logistic Regression	Decision Tree	Random Forest	Gaussian Naïve Bayes	XGB Classifier	Best score
Accuracy	0.78	0.79	0.80	0.66	0.83	XGB Classifier
Precision	0.75	0.81	0.81	0.70	0.83	XGB Classifier
Recall	0.83	0.76	0.78	0.55	0.83	XGB Classifier
F1 Score	0.79	0.78	0.80	0.62	0.83	XGB Classifier

To conclude, based on the above results, we chose Random Forest and XGB Classifier for further analysis.

We created a base model and tested the same on train and test dataset. Then, we proceeded with feature selection and hyper parameter tuning. Based on the results, we selected the best model that is XGB Classifier.

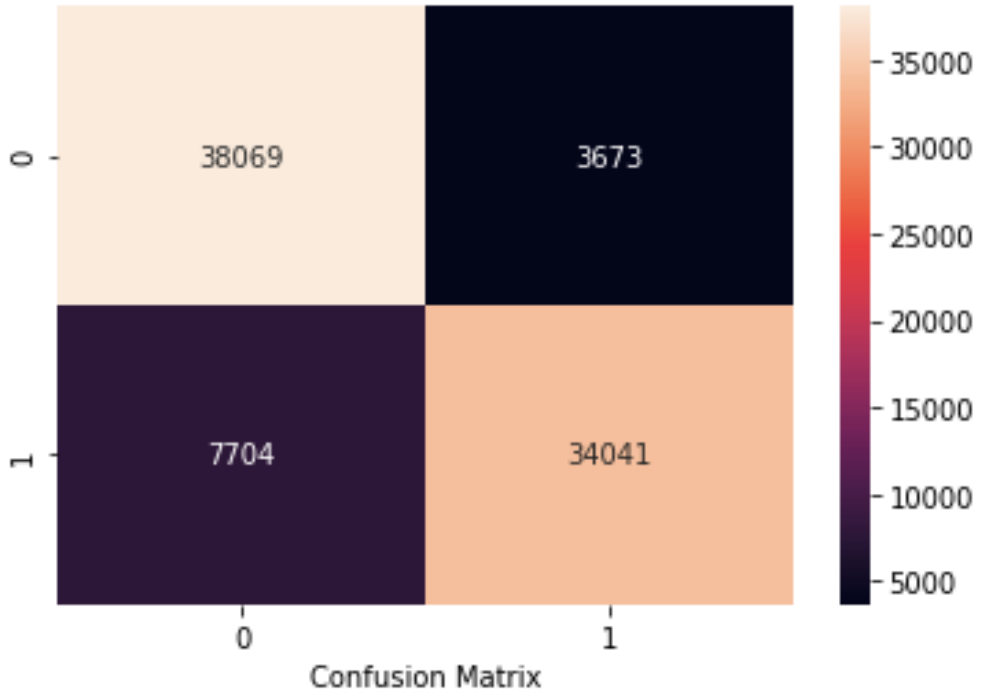
The following pages will talk about the corresponding results.

BASE MODEL - TRAIN

The classification report for train dataset for XGB classifier are as follows:

Train dataset report and confusion matrix

	Precision	Recall	F1 Score
0	0.83	0.91	0.87
1	0.9	0.82	0.86
Accuracy			0.86
Macro avg	0.87	0.86	0.86
Weighted avg	0.87	0.86	0.86



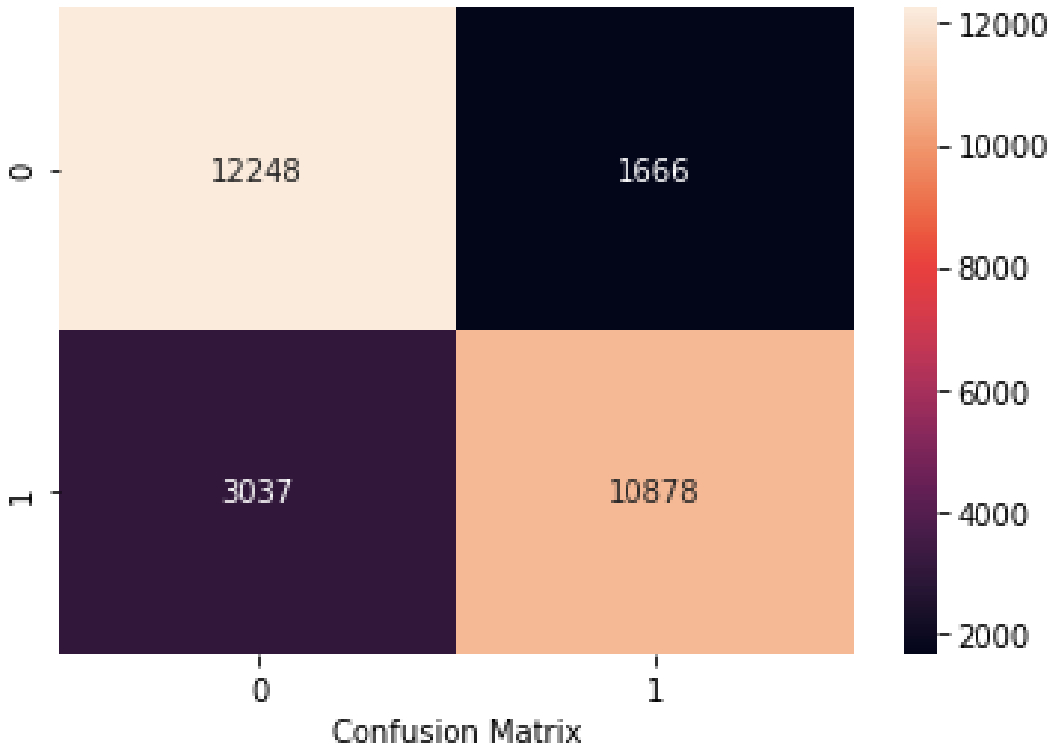
0: Issue closed with monetary relief
1: Issue closed without monetary relief

BASE MODEL - TEST

The classification report for test dataset for XGB classifier are as follows:

Test dataset report and confusion matrix

	Precision	Recall	F1 Score
0	0.8	0.88	0.84
1	0.87	0.78	0.82
Accuracy			0.83
Macro avg	0.83	0.83	0.83
Weighted avg	0.83	0.83	0.83



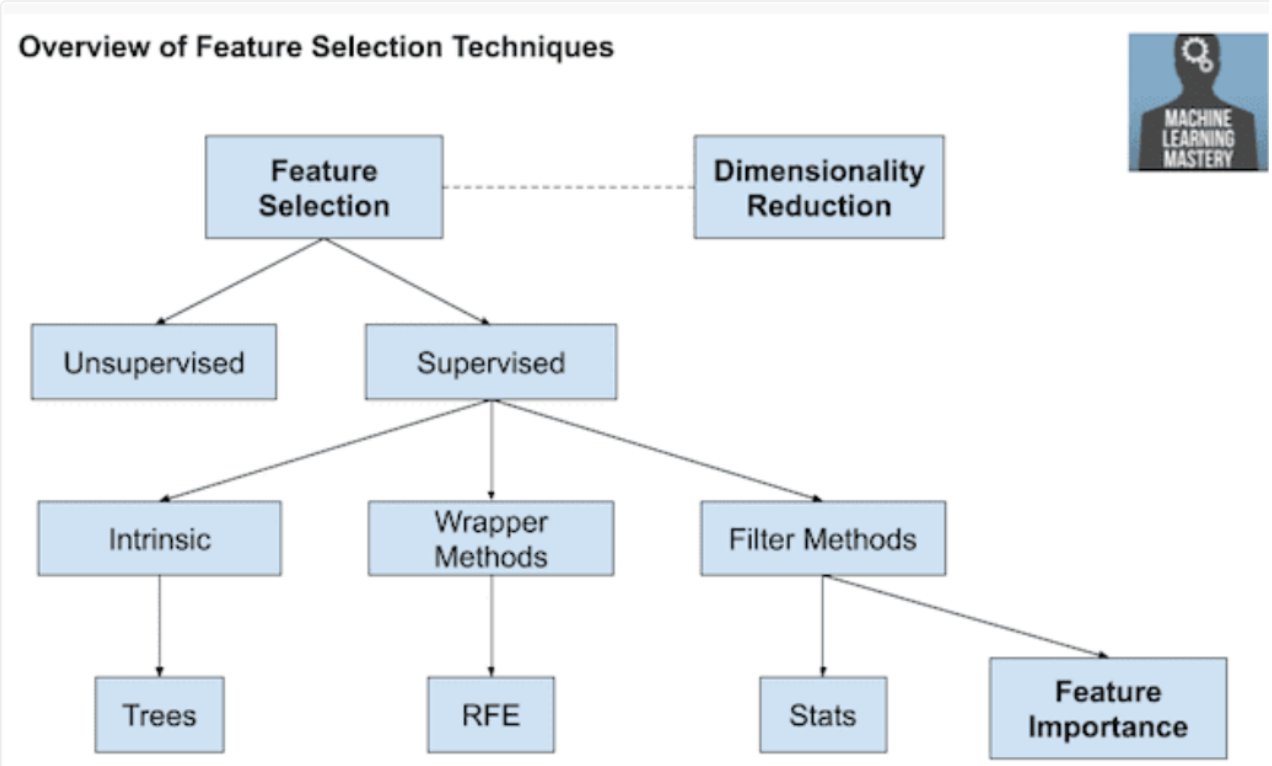
0: Issue closed with monetary relief
1: Issue closed without monetary relief

FEATURE SELECTION

The feature selection for supervised learning methods is mainly divided into wrapper methods and feature importances of which the former is employed during pre-model phase and the latter is for giving insights to the business post the model building.

We used forward feature selection to identify the best features and they are as follows; Product, Sub-product, Issue, Company, Submitted via and Timely response.

We then proceeded with hyper-parameter tuning for the chosen features.

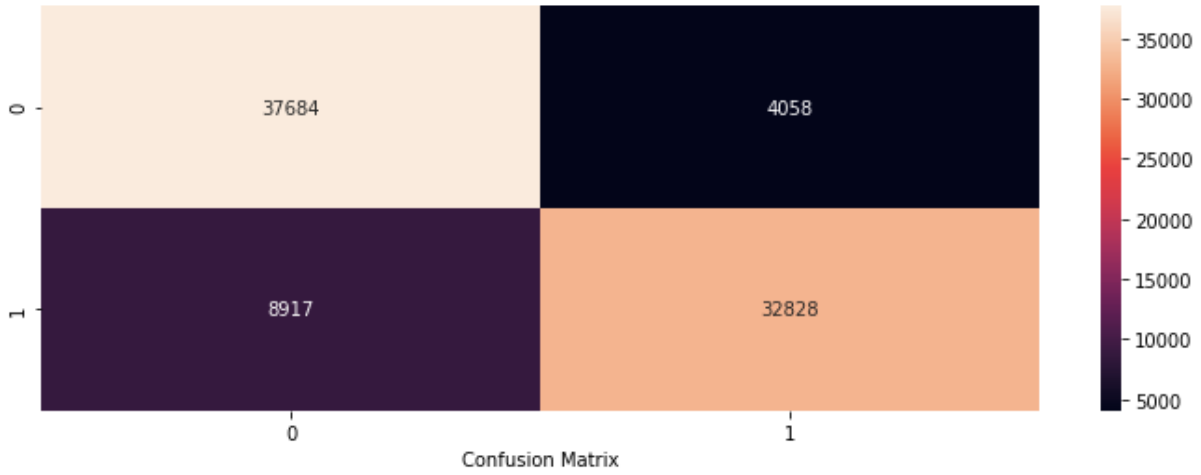


GRIDSEARCH CV

For GridSearch CV, we performed recursive parameter finding across multiple iterations. The best parameters obtained are as follows: Learning_rate =0.2, n_estimators=500, max_depth=5, min_child_weight = 4, gamma =0.2, subsample = 0.75 and colsample_bytree =0.55. We built our model based on these results.

Train dataset report and confusion matrix

	Precision	Recall	F1 Score
0	0.81	0.9	0.85
1	0.89	0.79	0.83
Accuracy			0.84
Macro avg	0.85	0.84	0.84
Weighted avg	0.85	0.84	0.84

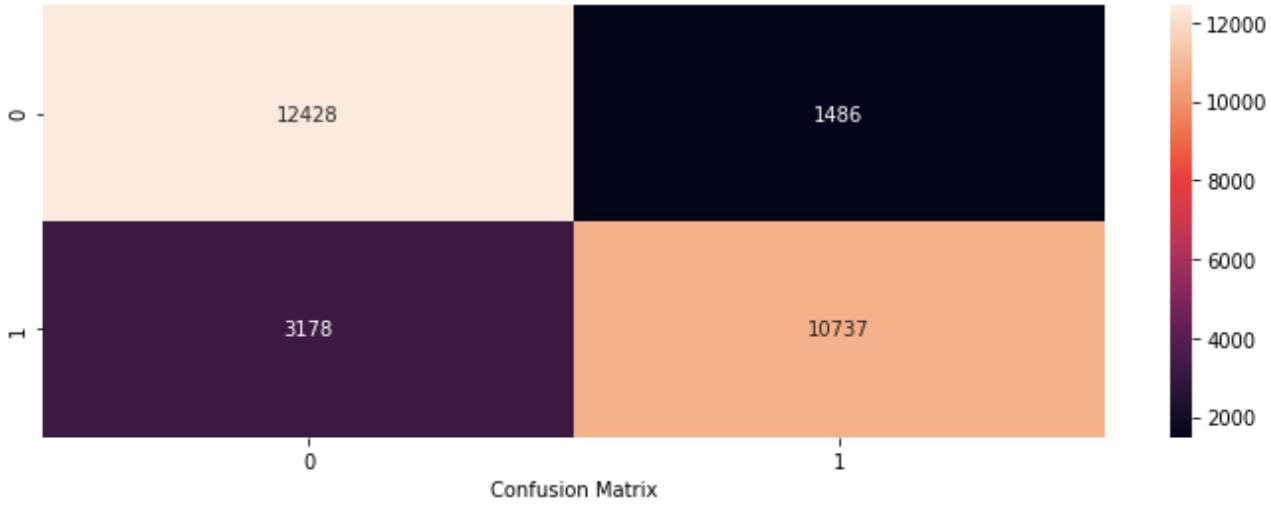


0: Issue closed with monetary relief
1: Issue closed without monetary relief

GRIDSEARCH CV

Test dataset report and confusion matrix

	Precision	Recall	F1 Score
0	0.8	0.89	0.84
1	0.88	0.77	0.82
Accuracy			0.83
Macro avg	0.84	0.83	0.83
Weighted avg	0.84	0.83	0.83



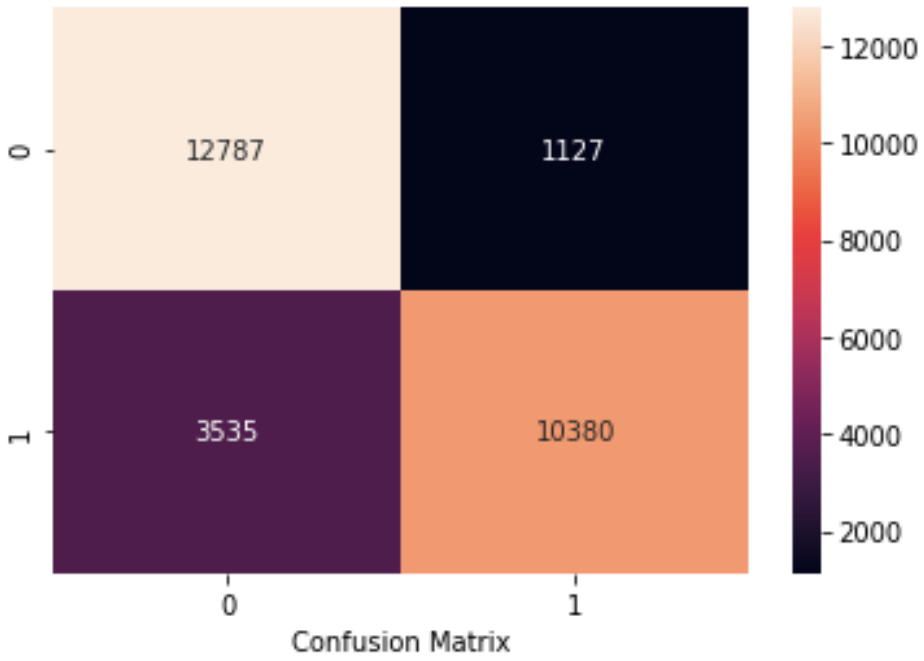
0: Issue closed with monetary relief
1: Issue closed without monetary relief

THRESHOLD TUNING

Our objective here is to reduce and predict issues with monetary relief correctly as and when the issues arise. If otherwise, we will believe the issue isn't that significant and will not cause any monetary loss to us.

Keeping our objective in mind, after building our final model we proceeded with threshold tuning. In order to reduce false positives, we adjusted the threshold value to 0.43 to bring about the desired results. The alpha error is around 8% and beta error is 25.40%.

Upon doing so, our confusion matrix for test report is as follows:

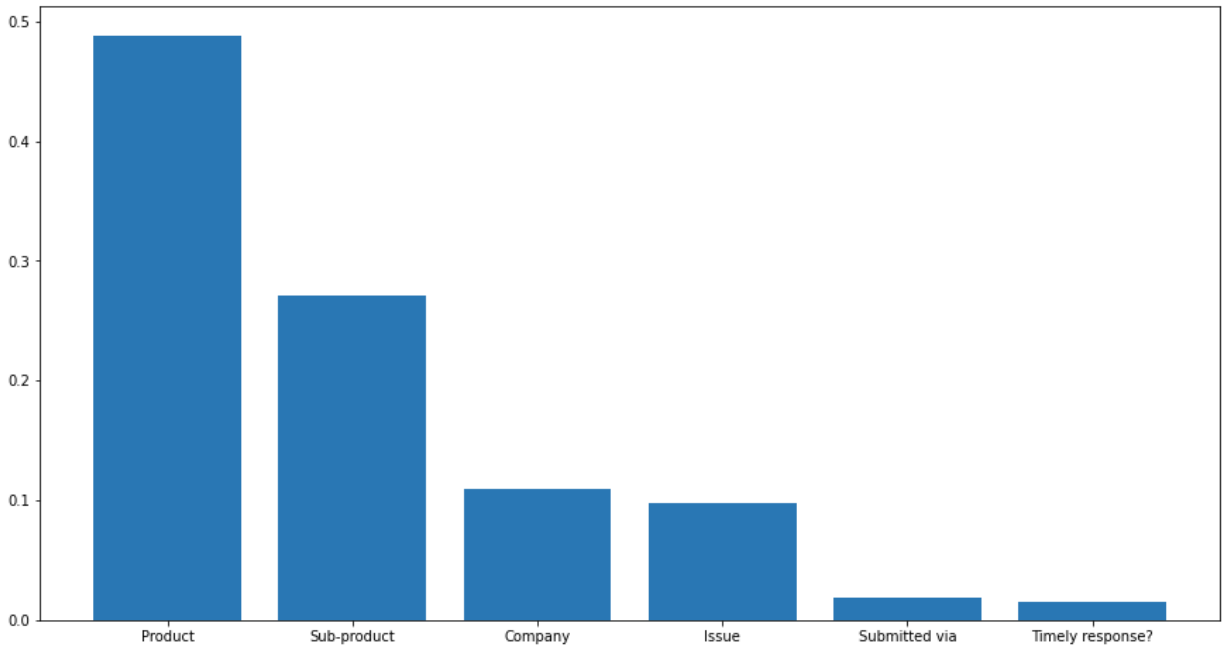


0: Issue closed with monetary relief
1: Issue closed without monetary relief

FEATURE IMPORTANCE

After building our final model, we performed 'feature importances' to get the most significant features.

	Features	Importance
0	Product	0.488457
1	Sub-product	0.271303
3	Company	0.108958
2	Issue	0.097758
4	Submitted via	0.018782
5	Timely response?	0.014742



OBJECTIVE II

Our objective is to predict if there will be timely responses for the issue, so that respected team can emphasize more on such complaints to resolve it on time.

Our target column **'Timely response'** had unequal proportion of the levels. 'Yes' was of 97.49% whereas 'NO' was just 2.51%.

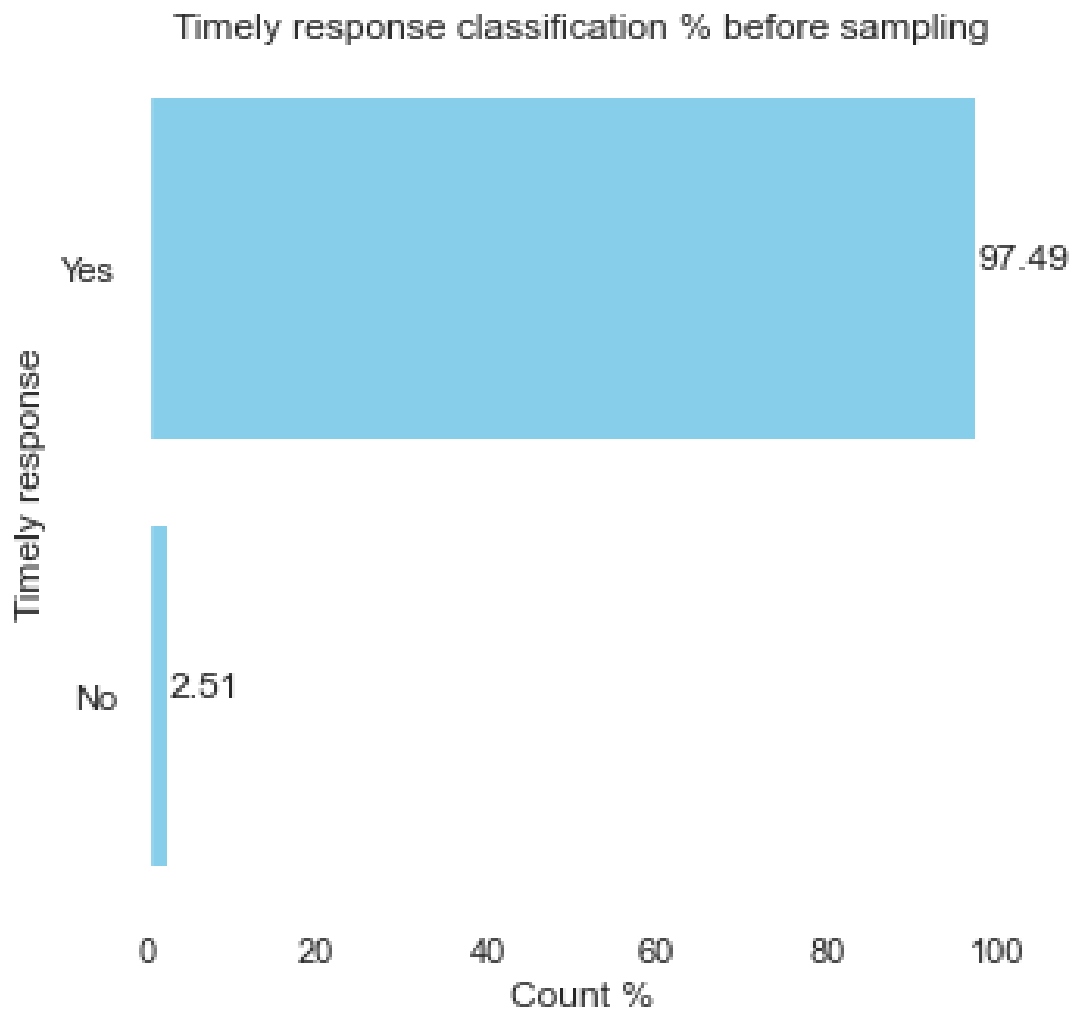


Figure 7: Before applying random sampling technique

To overcome this issue, we applied random sampling technique and selected equal number of rows in each column.

Our target column now has equal representation of each level.

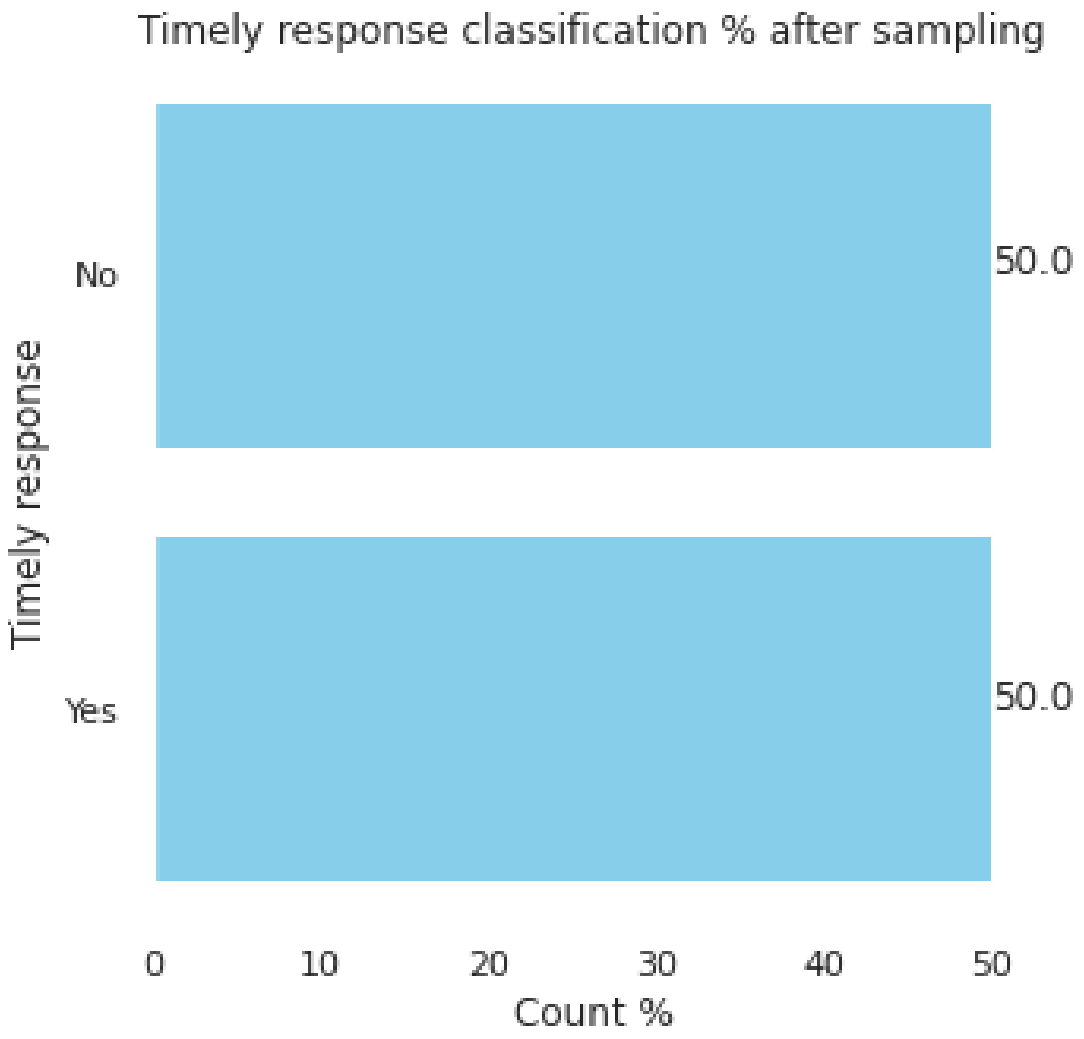


Figure 8: After applying random sampling technique

PREPROCESSING

For our objective, we created a data subset using random sampling technique.

After that, our target columns had equal representation of the levels.

The final dataset contains 54446 rows and 14 columns. It was prepared for further analysis by removing or modifying the data that has been duplicated, incorrect, incomplete or irrelevant.

NULL IMPUTATION

Null values have been treated accordingly with respect to the null count in the column;

Column Names	Missing (%)	Handling Method
Sub-product	10.88	Conditional Imputing
Sub-issue	31.24	Column Dropped
State	1.64	Mode Imputation
Consumer consent provided?	32.30	Column Dropped
Consumer disputed?	44.34	Column Dropped

FEATURE ENGINEERING

1. 'State' column had 62 entries which was then clubbed based on the US map into regions having 6 levels.

2. '**Duration'** column had been created by subtracting **'Date sent to company'** and **'Date received**'. This columns basically talks about the time taken for the filed response to reach the company.

3. Post which columns such as 'State', 'Zip Code', 'Date sent to company' and 'Date received' were dropped.

4. The identifier column ('Complaint ID') was dropped from further analysis.

EDA

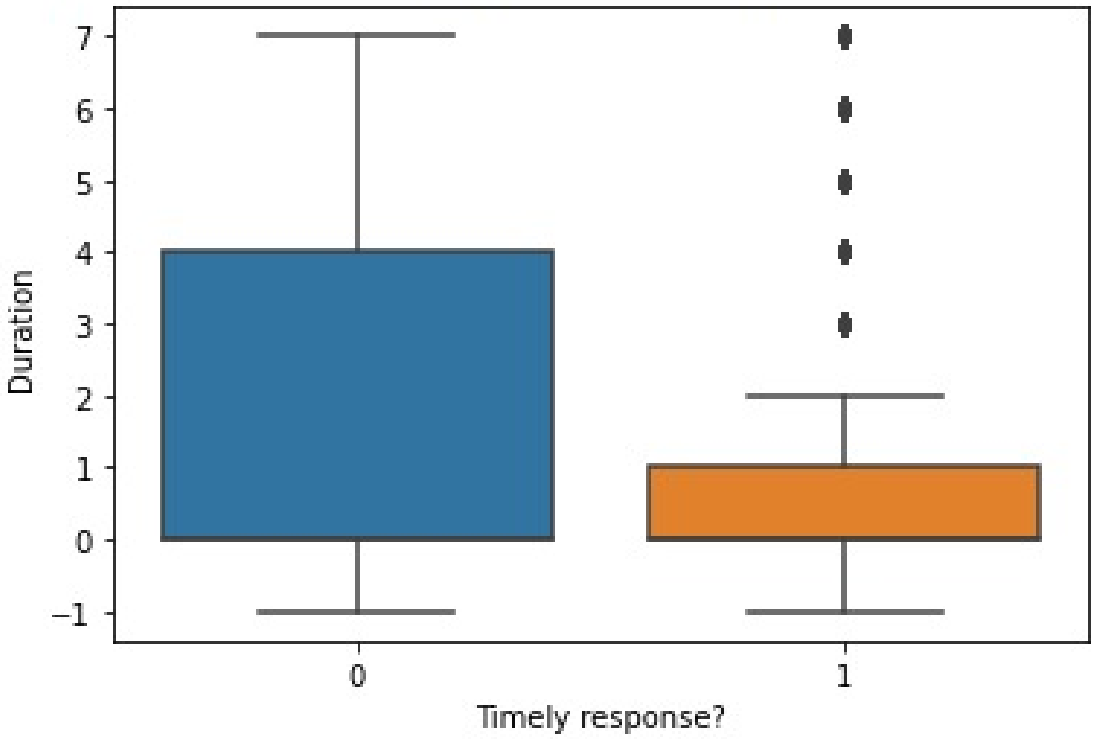


Figure 9: Boxplot of duration with the target column

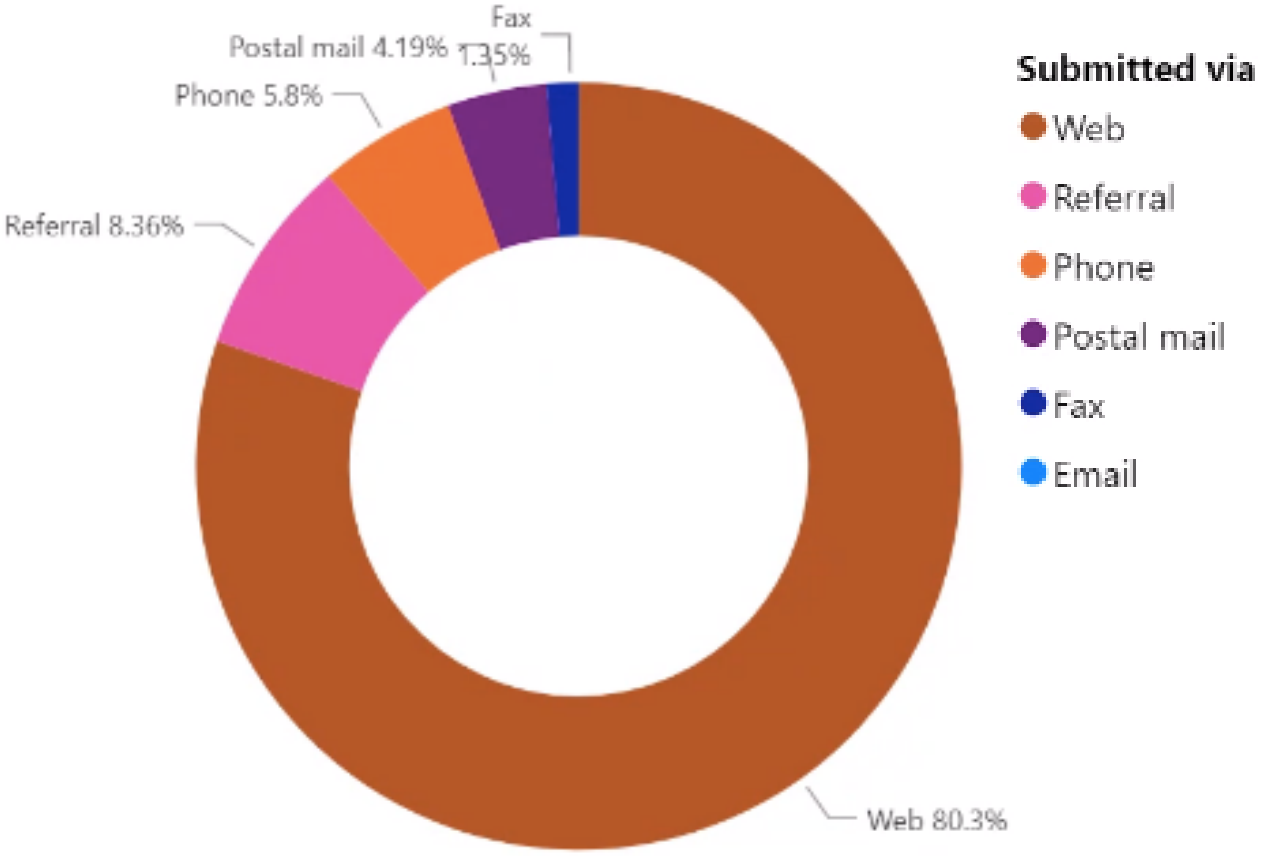


Figure 10: Mode of filing complaints

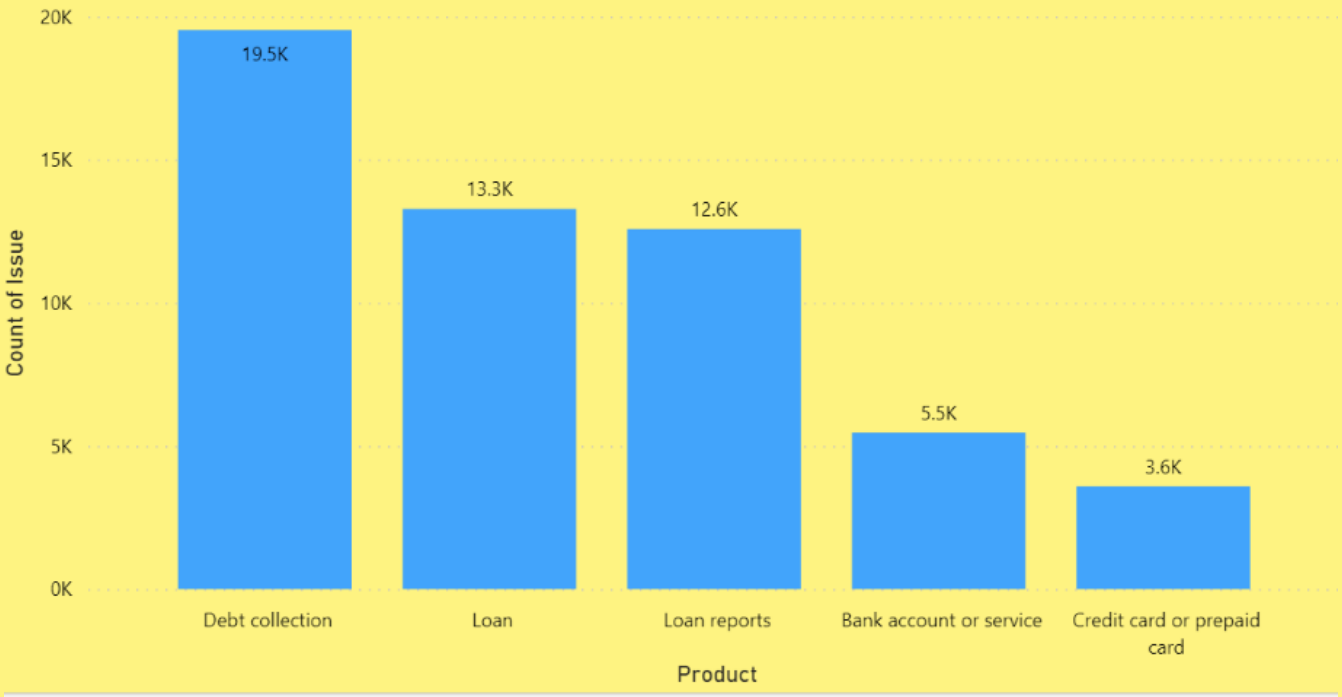


Figure 11: Count of issue by product

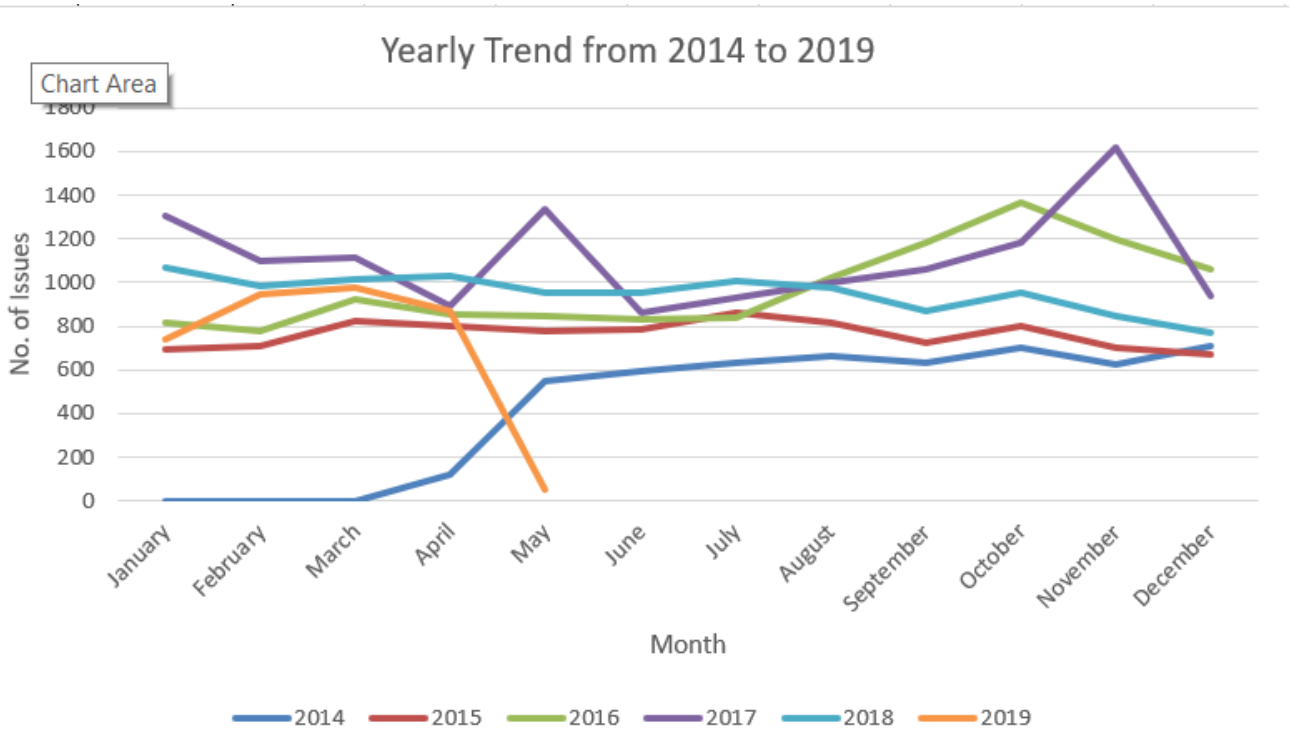


Figure 12: Count of issues yearly trend

ENCODING

Each categorically column was encoded with suitable technique as follows:

Column Names	No. of unique values	Technique
Product	5	Factorize Encoding
Sub-Product	53	Frequency Encoding
Issue	94	Frequency Encoding
Company	3385	Frequency Encoding
Submitted via	6	Factorize Encoding
Timely response	2	Factorize Encoding
Company response to consumer	6	Factorize Encoding
Region	6	Factorize Encoding

MODEL BUILDING

We trained our dataset in various classifiers and evaluated the model based on performance metrics such as:

- Accuracy score
- Recall score
- Precision score
- F1 score

	Logistic Regression	Decision Tree	Random Forest	Gaussian Naïve Bayes	XGB Classifier	Best score
Accuracy	0.61	0.78	0.79	0.63	0.83	XGB Classifier
Precision	0.61	0.77	0.79	0.58	0.84	XGB Classifier
Recall	0.63	0.80	0.80	0.92	0.83	Gaussian Naïve Bayes
F1 Score	0.62	0.78	0.79	0.71	0.83	XGB Classifier

To conclude, based on the above results, we chose Random Forest and XGB Classifier for further analysis.

We created a base model and tested the same on train and test dataset. Then, we proceeded with feature selection and hyper parameter tuning. Based on the results, we selected the best model that is XGB Classifier.

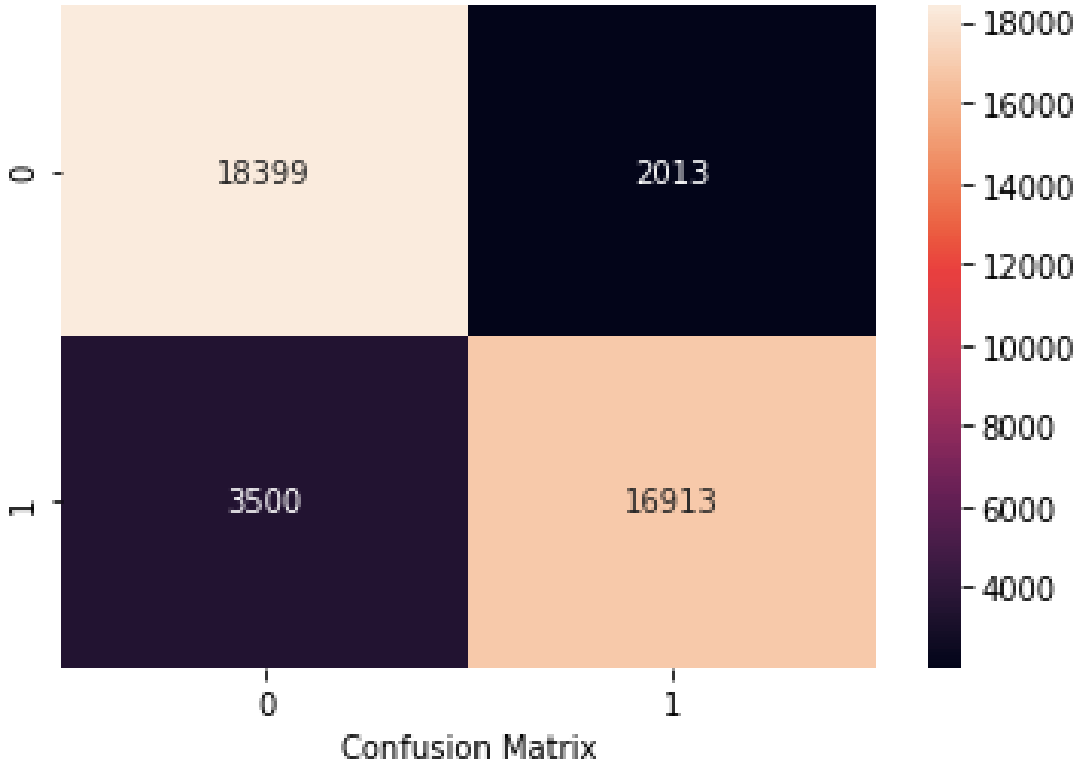
The following pages will talk about the corresponding results.

BASE MODEL - TRAIN

The classification report for train dataset for XGB Classifier are as follows:

Train dataset report and confusion matrix

	Precision	Recall	F1 Score
0	0.84	0.9	0.87
1	0.89	0.83	0.86
Accuracy			0.86
Macro avg	0.87	0.86	0.86
Weighted avg	0.87	0.86	0.86



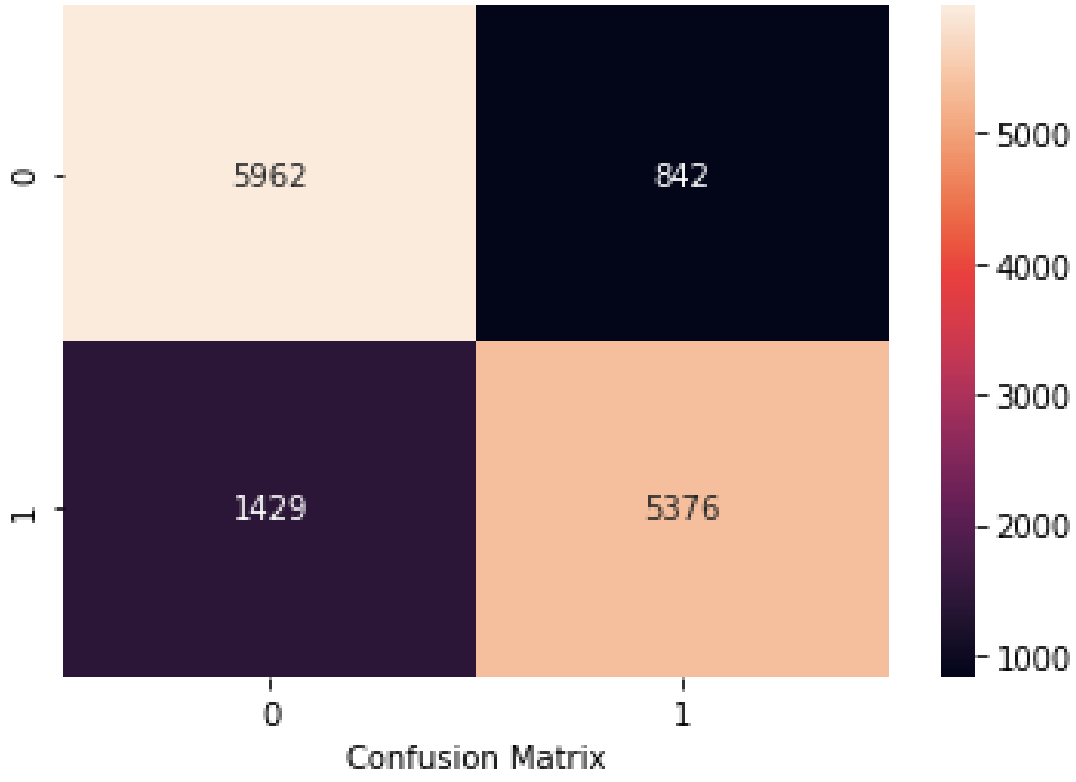
- 0: Timely response -No
- 1: Timely response - Yes

BASE MODEL - TEST

The classification report for test dataset for XGB classifier are as follows:

Test dataset report and confusion matrix

	Precision	Recall	F1 Score
0	0.81	0.88	0.84
1	0.86	0.79	0.83
Accuracy			0.83
Macro avg	0.84	0.83	0.83
Weighted avg	0.84	0.83	0.83



- 0: Timely response -No
- 1: Timely response - Yes

FEATURE SELECTION

We used forward feature selection to identify the best features. For this purpose, we had to import SequentialFeatureSelector from mlxtend.feature_selection. With n_jobs=-1, it uses all the cores for parallel computing. Further we fixed k_features ='best' and scoring as F1.

```
from xgboost import XGBClassifier
from mlxtend.feature_selection import SequentialFeatureSelector
forward_feature_selector = SequentialFeatureSelector(XGBClassifier(n_jobs=-1),
           k_features='best',
           forward=True,
           verbose=2,
           scoring='f1',
           cv=5)

fselector = forward_feature_selector.fit(X_train, y_train)
```

```
fselector.k_feature_names_
```

```
('Product', 'Sub-product', 'Company', 'Regions')
```

The best features are as follows; Product, Sub-product, Company and Regions.

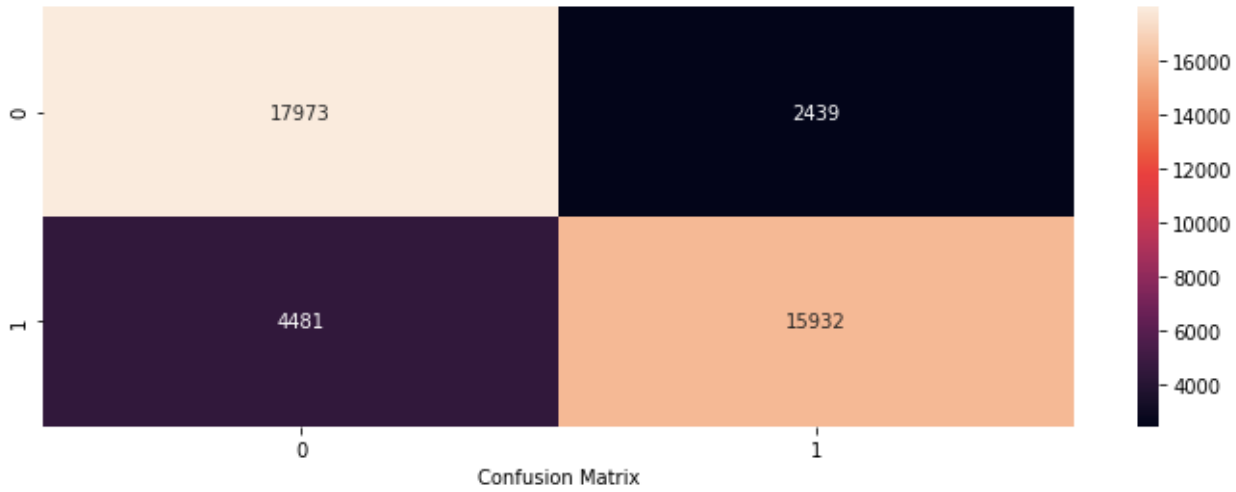
We then proceeded with hyper-parameter tuning for the chosen features.

GRIDSEARCH CV

For GridSearch CV, we performed recursive parameter finding across multiple iterations. The best parameters obtained are as follows: Learning_rate =0.1, n_estimators=500, max_depth=4, min_child_weight = 5, gamma =0.2, subsample = 0.80 and colsample_bytree =0.75. We built our model based on these results.

Train dataset report and confusion matrix

	Precision	Recall	F1 Score
0	0.8	0.88	0.84
1	0.87	0.78	0.82
Accuracy			0.83
Macro avg	0.83	0.83	0.83
Weighted avg	0.83	0.83	0.83

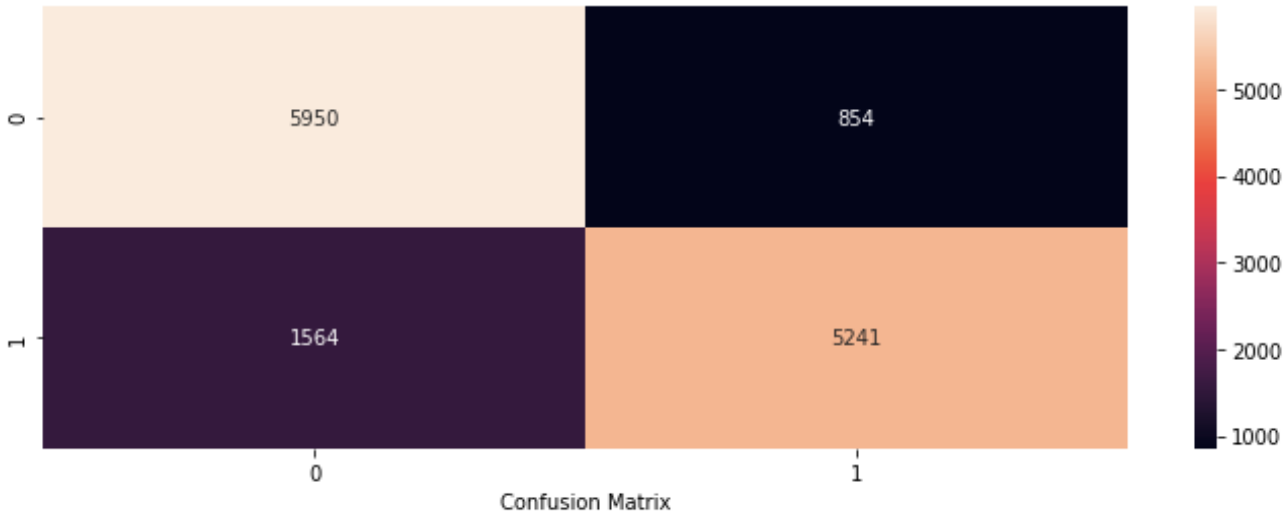


- 0: Timely response -No
- 1: Timely response - Yes

GRIDSEARCH CV

Test dataset report and confusion matrix

	Precision	Recall	F1 Score
0	0.79	0.87	0.83
1	0.86	0.7	0.81
Accuracy			0.82
Macro avg	0.83	0.82	0.82
Weighted avg	0.83	0.82	0.82



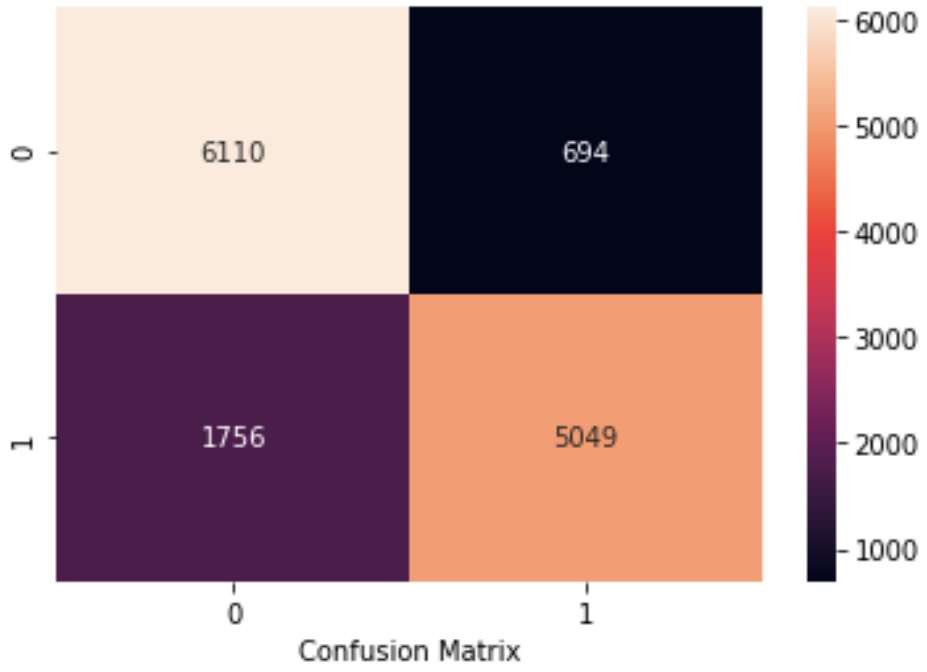
- 0: Timely response -No
- 1: Timely response - Yes

THRESHOLD TUNING

Our objective here is to reduce and predict issues that requires prompt timely response so they are responded within a given time frame. If otherwise, we will believe the issue isn't significant and will not provide the required assistance for the same.

Keeping our objective in mind, after building our final model we proceeded with threshold tuning. In order to reduce false positive, we adjusted the threshold value to 0.40 to bring about the desired results. The alpha error is around 10% and beta error is 25.80%.

Upon doing so, our confusion matrix for test report is as follows:

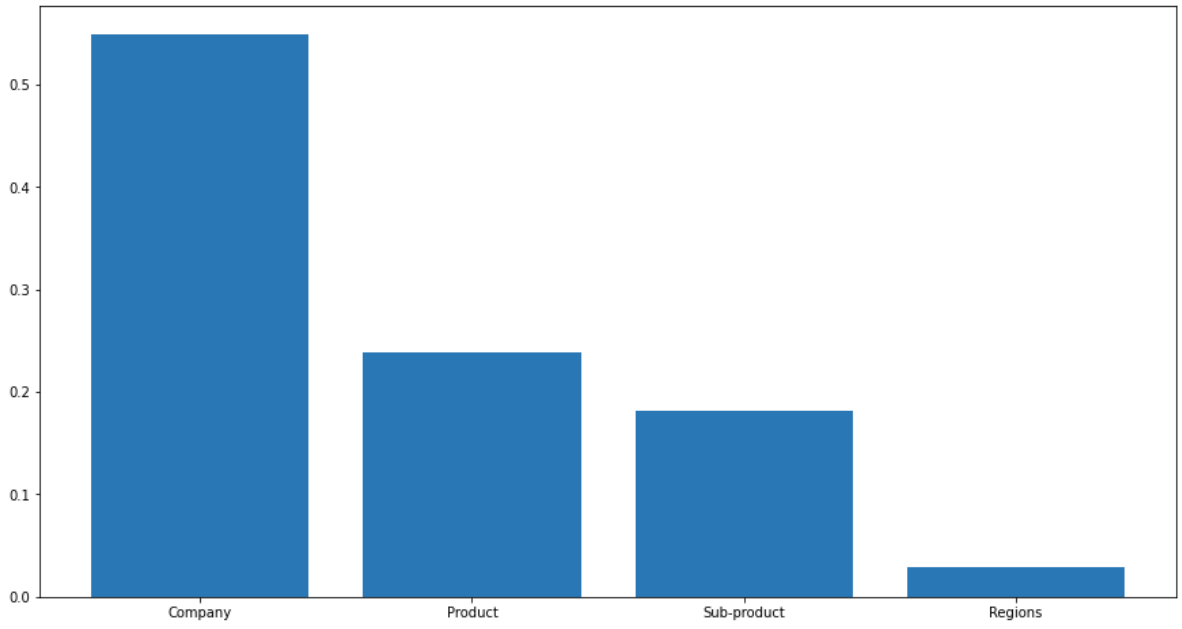


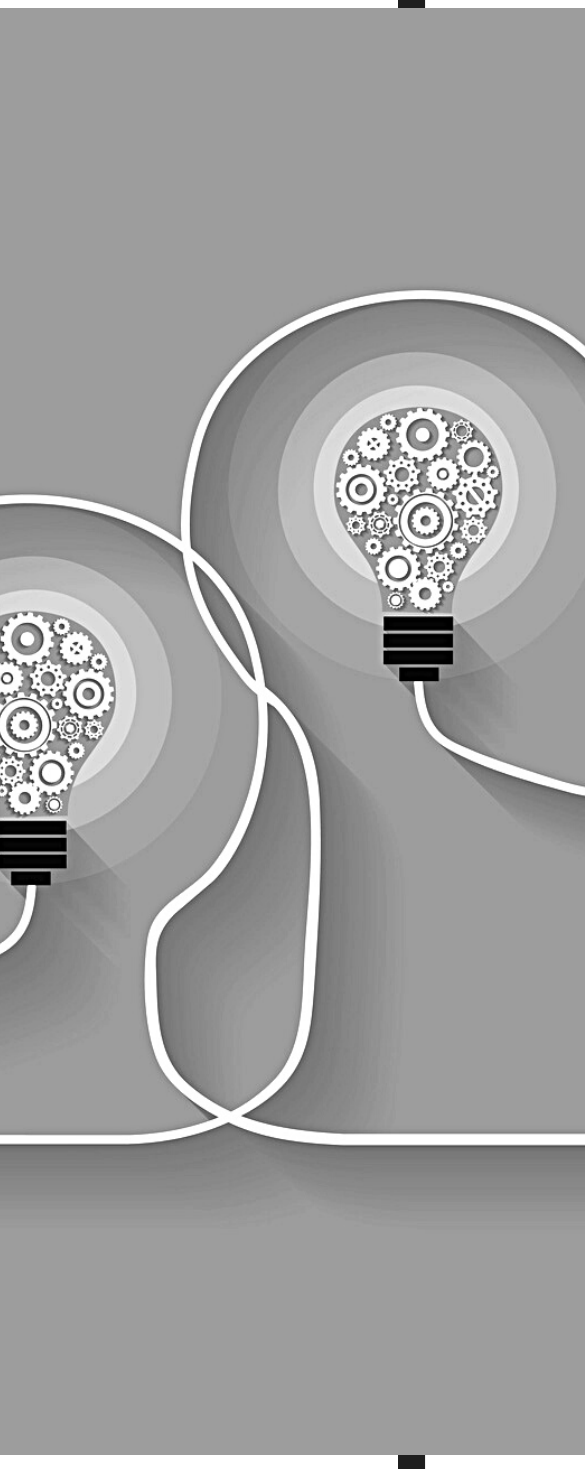
- 0: Timely response -No
- 1: Timely response - Yes

FEATURE IMPORTANCE

After building our final model, we performed 'feature importances' to get the most significant features.

	Features	Importance
2	Company	0.549823
0	Product	0.239261
1	Sub-product	0.181707
3	Regions	0.029209





BUSINESS INSIGHTS

"Credit card or report card", the product constituted more number of issues with monetary reliefs while "Debt collection" accounted for more number of issues with no on-time response. Despite having a large volume of issues, 'Debt collection' constituted low number of issues that closed with monetary relief.

"Account opening, Closing or Management" is found to be a cause of major concern for issues with monetary relief whereas "Incorrect information on credit report" accounted for issues that has no on-time response.

"Bank of America, National Association" received most number of complaints for objective I followed by Citibank N.A. and JPMorgan Chase and co. whereas "Equifax, Inc." received most number of complaints for objective II followed by Wells Fargo & company and Experian Information Solutions Inc.

CONCLUSION

01

Building Final model

- Built a final model after performing hyper-parameter tuning to analyse and derive business insights.
- Model had an accuracy score of over 80% for both the objectives,

02

Proposition

- Our model will help the company to identify issues that require monetary relief. Timely action on the same will reduce the loss.
- Allows company to identify the department with most issues, so as to take quick action and to make effective changes.

03

Future scope

- To perform computational linguistics on customer reviews.
- To build deep learning classification models for improved accuracy score.