AirBnB

Group 10

Arjun Chanda

Ateeksha Chaudhary

Garrett Chaffey

Yuexi Li

Qiuhan Li

Introduction

Discover Amazing Places

Find hosts with extra rooms, entire homes, and unique accommodations like castles and igloos.

Book a Stay

Connect with hosts, confirm travel dates, and pay—all through Airbnb's trusted services.

Travel

Feel at home, anywhere you go in the world. Learn more about how to travel on Airbnb.

Learn more about hosting on Airbnb and show your guests they're right where they belong.

Company Overview

- $3.4 billion in revenue in 2020
- 193 million bookings in Airbnb in 2020
- Over 7 million listings in Airbnb, run by 4 million hosts

Objective

Predict if a property will receive a good or bad rating based on sentiment analysis and average rating score.

Value Proposition of Project

Small Airbnb Managers:

- Determine most important attributes
- Compare property to competition

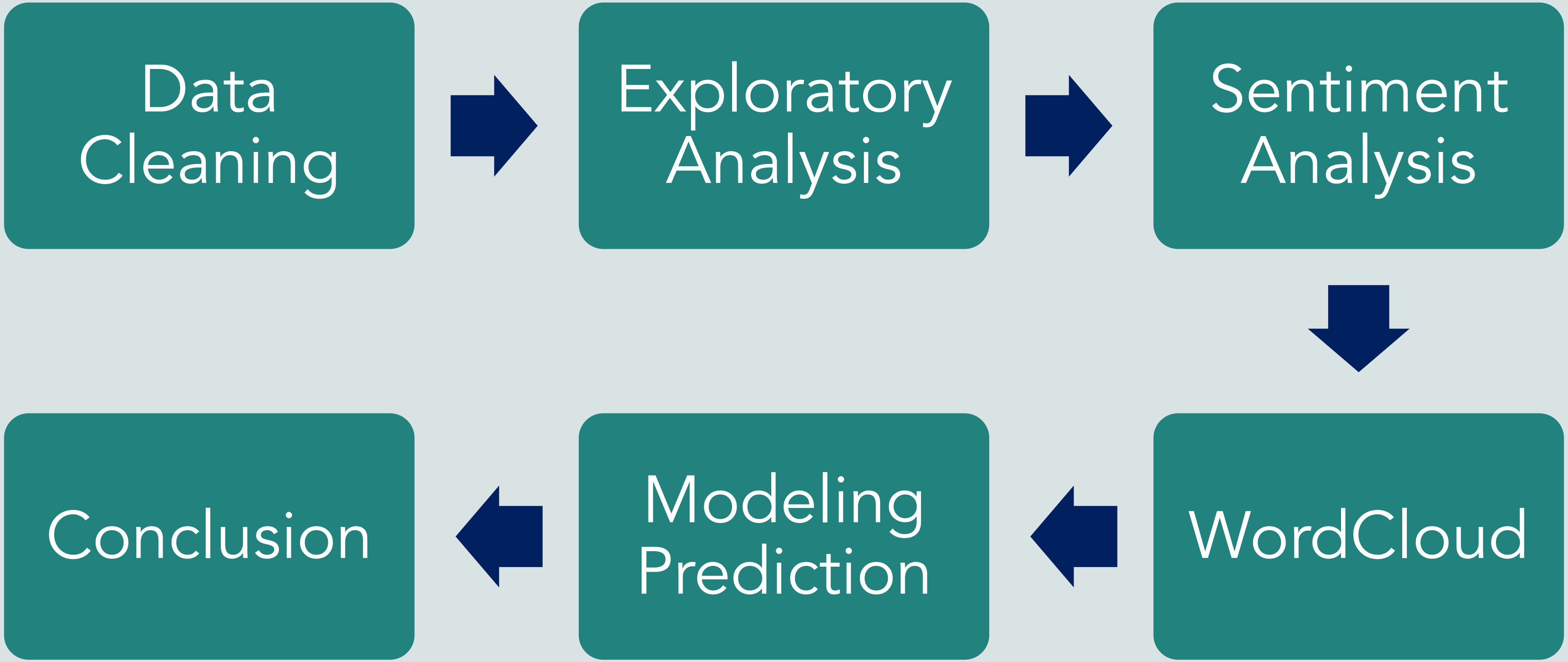
Large Airbnb Managers:

- Clarify user rating drivers
- Guide purchasing decisions

Airbnb Corporate:

- Vet aspiring Airbnb hosts
- Improve recommendation algorithm

Process Overview



Data Cleaning

Exploratory Analysis

Sentiment Analysis

WordCloud

Modeling Prediction

Conclusion

Data Cleaning

Original Dataset

- Two datasets - ‘Listings’ and ‘Reviews’
    - Combined into one
    - Reviews - 876,200 entries
    - Listings - 37,713 properties

- 43+ columns/attributes in combined dataset

- Reviews in multiple languages

- “Corrupted” reviews containing non-character symbols like ™, ®, ℵ etc

Dataset Pre-processing

1. Remove all non strings, digits and symbols

2. Remove punctuation, white spaces and converting upper case to lower case

3. Remove Stopwords using nltk stopwords

4. Remove null values

Cleaned Dataset

- 15,722 rows and two columns in ‘Reviews’
- 12,876 properties in ‘Listings’
- 42 columns in ‘Listings’

10 variables included from 2 different datasets, namely Listings and Reviews, such as:

Cleanliness

Subjectivity

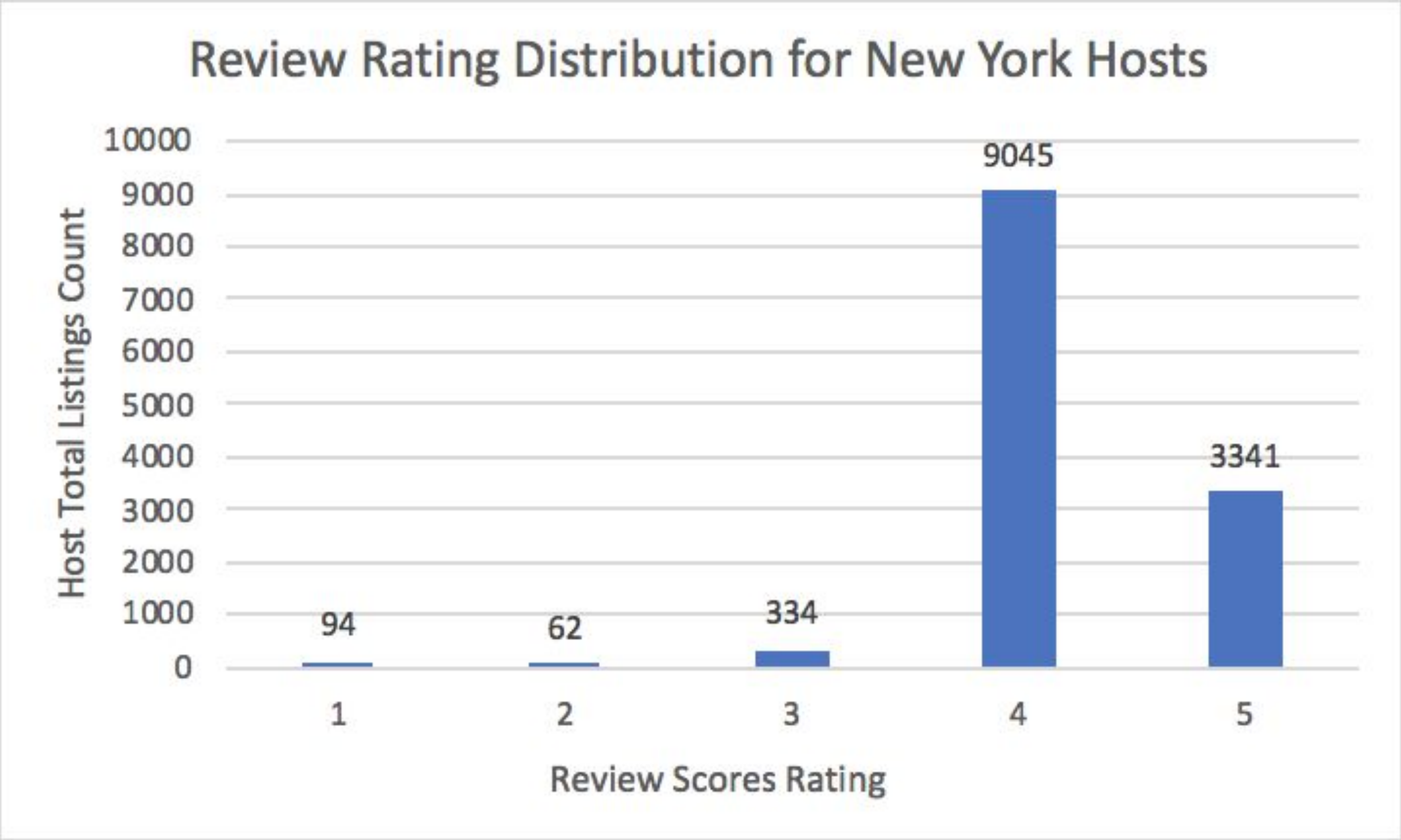
Polarity

Value

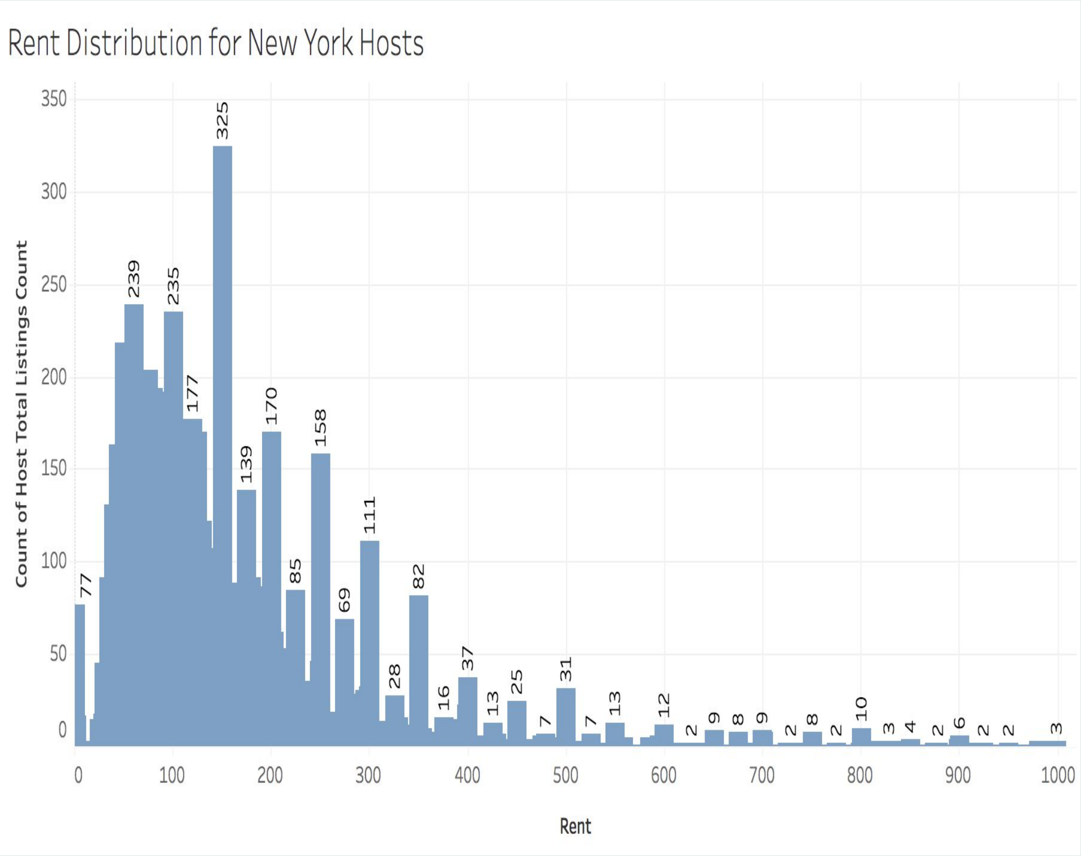
Rent

Exploratory Analysis

Demographic Visualizations

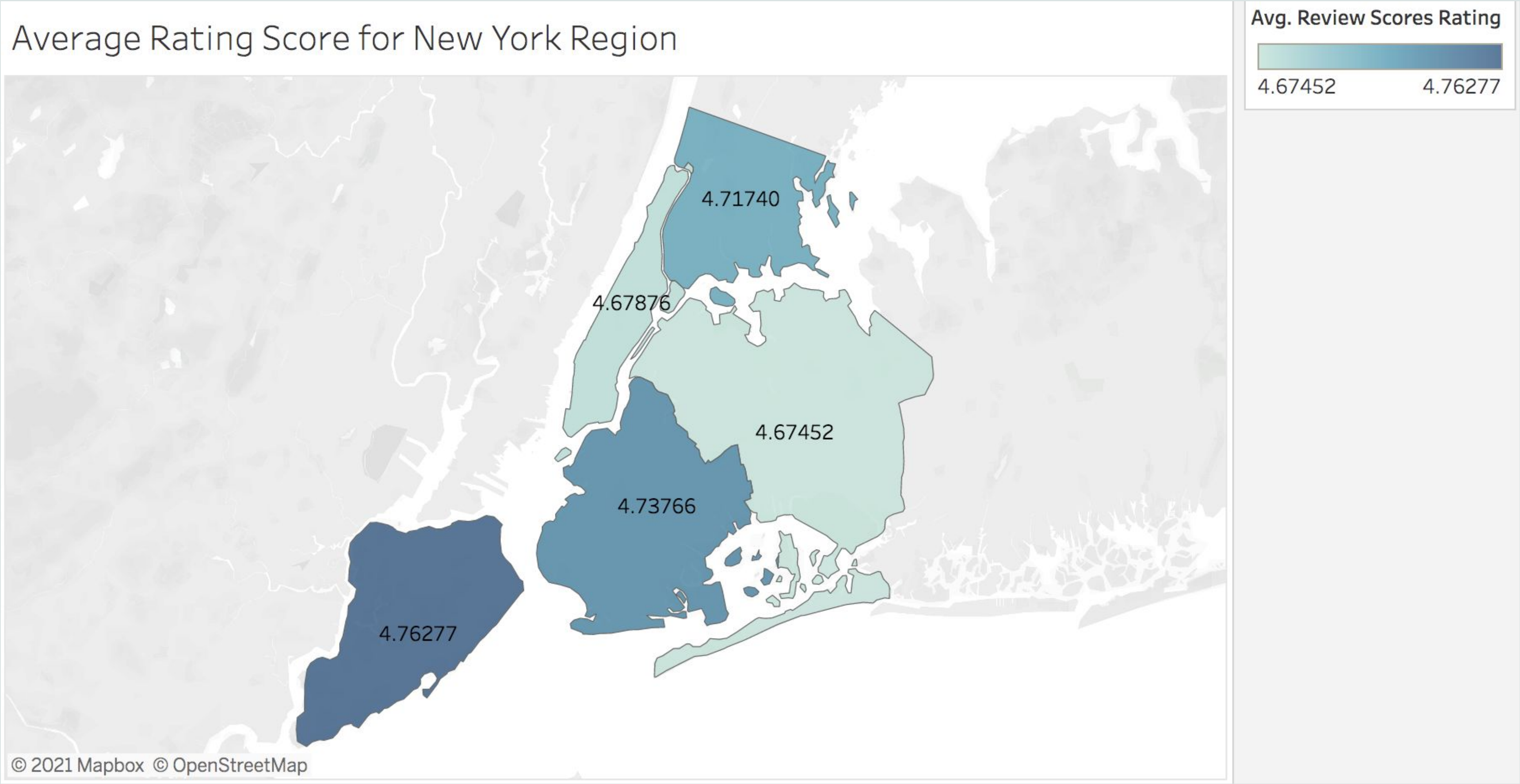


Review Rating Distribution for New York Hosts

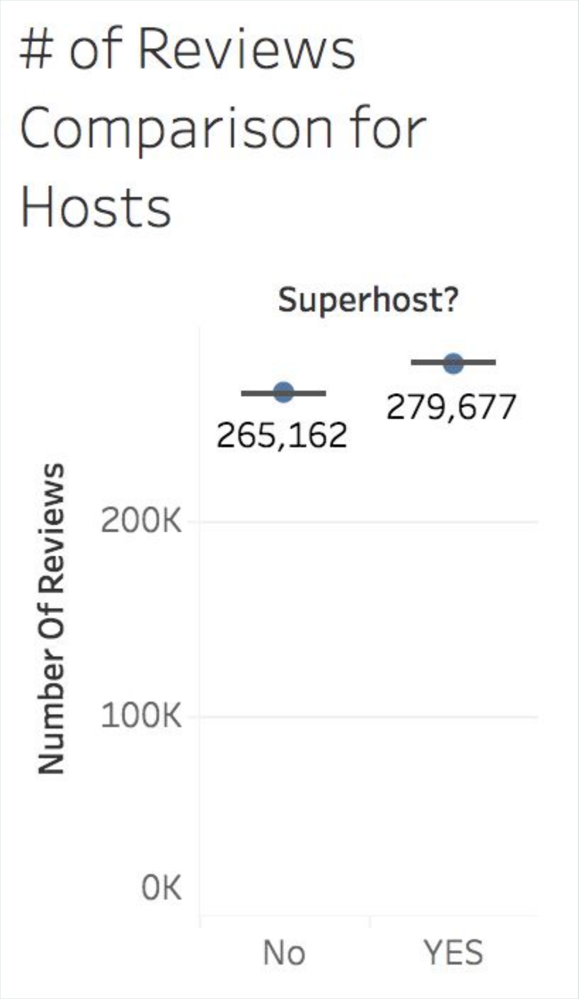


Rent Distribution for New York Hosts

Demographic Visualizations

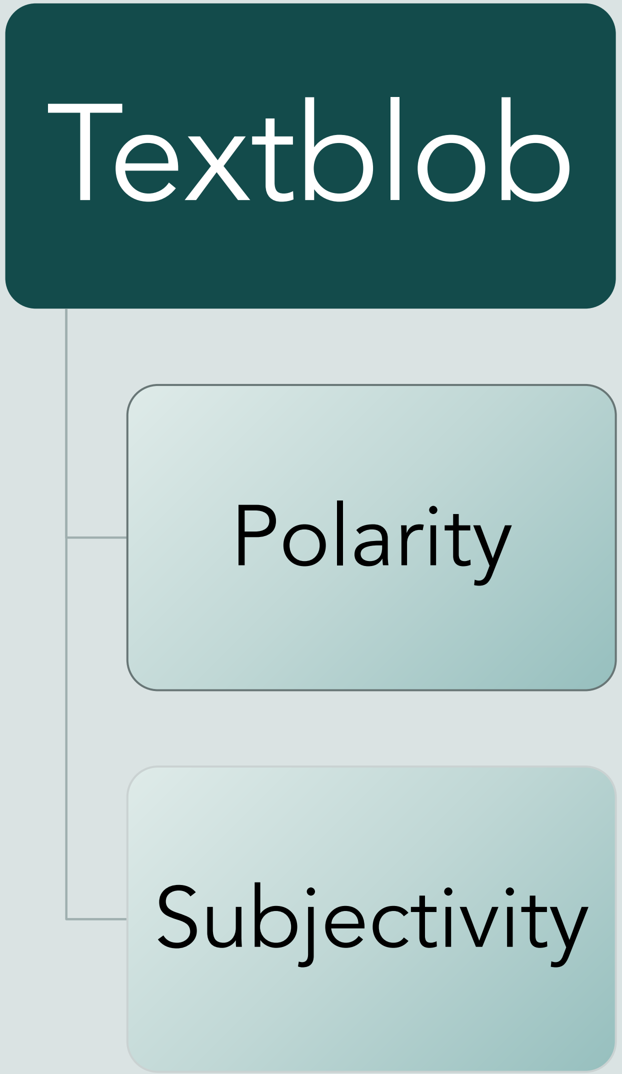


Average Rating Score for New York Region



# of Reviews Comparison for Hosts

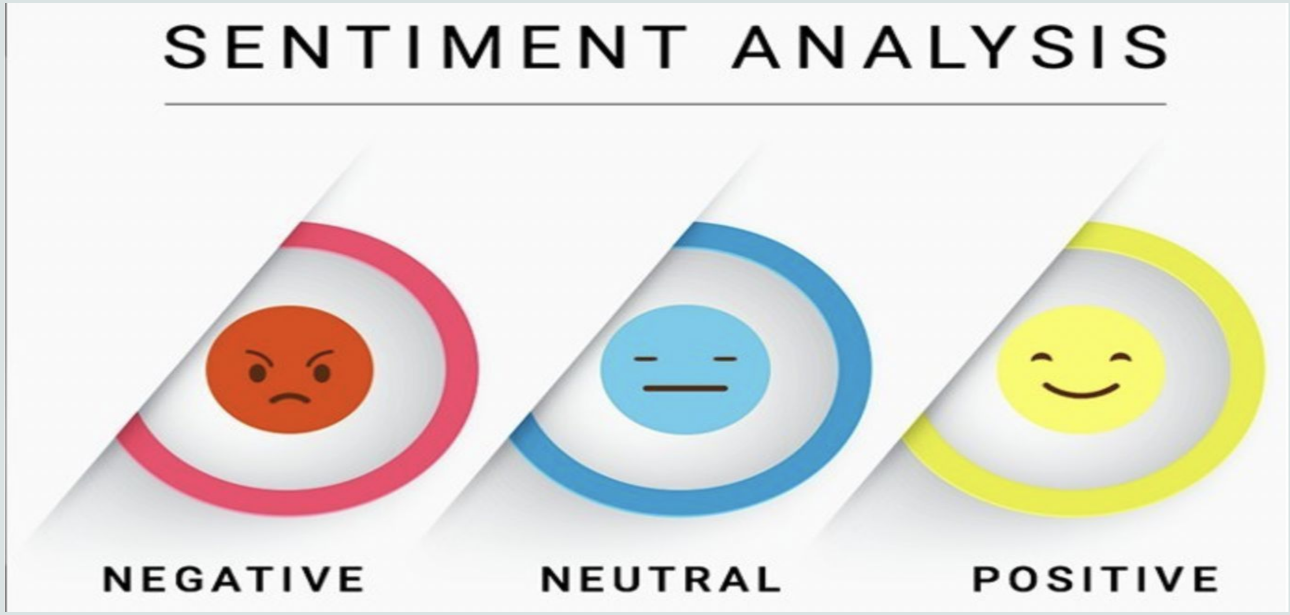
Sentiment Analysis



Textblob

Polarity

Subjectivity



Negative Neutral Positive

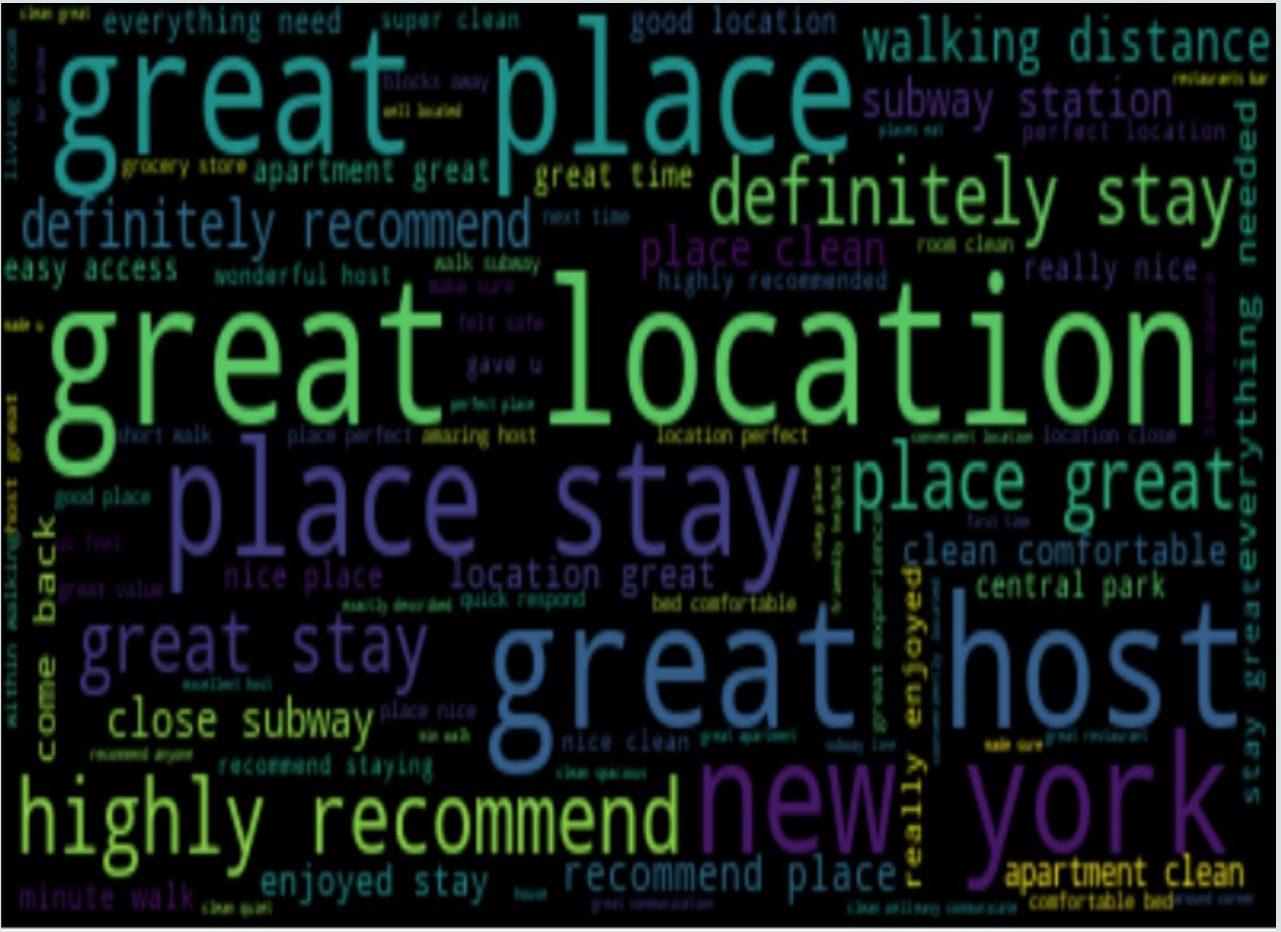
-1 0 1

Objective (factual) Subjective (opinion)

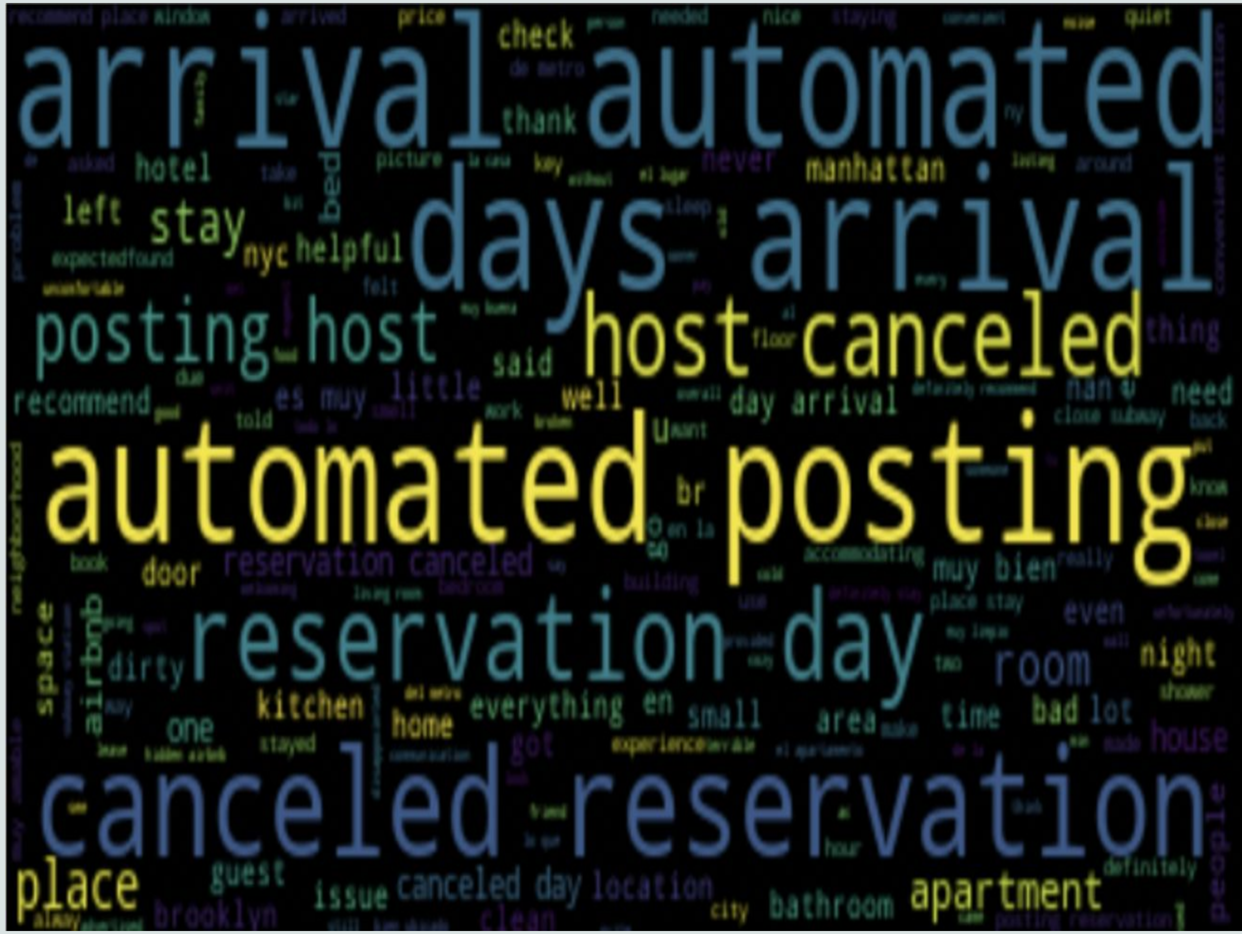
0 1

WordCloud

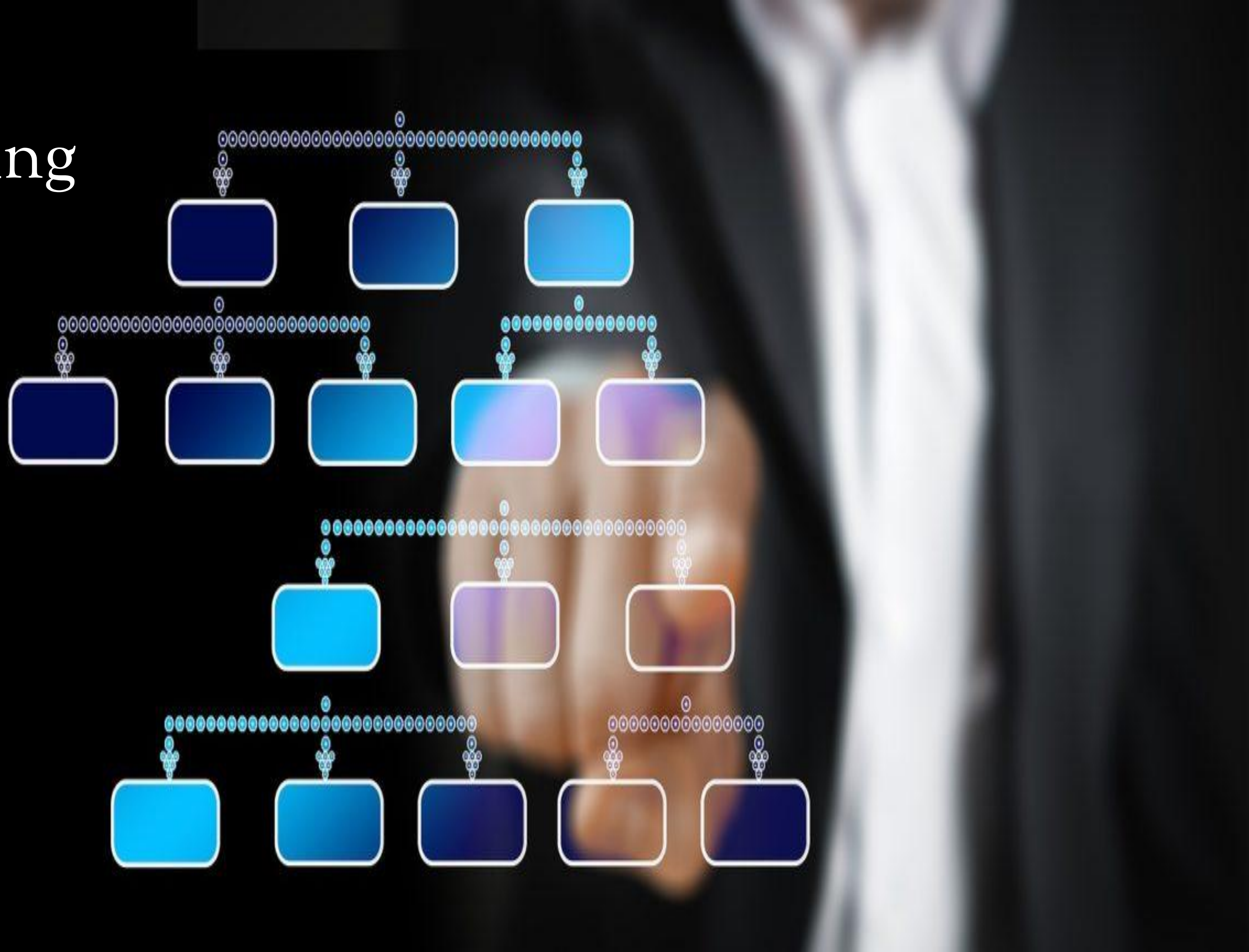
Positive



Negative

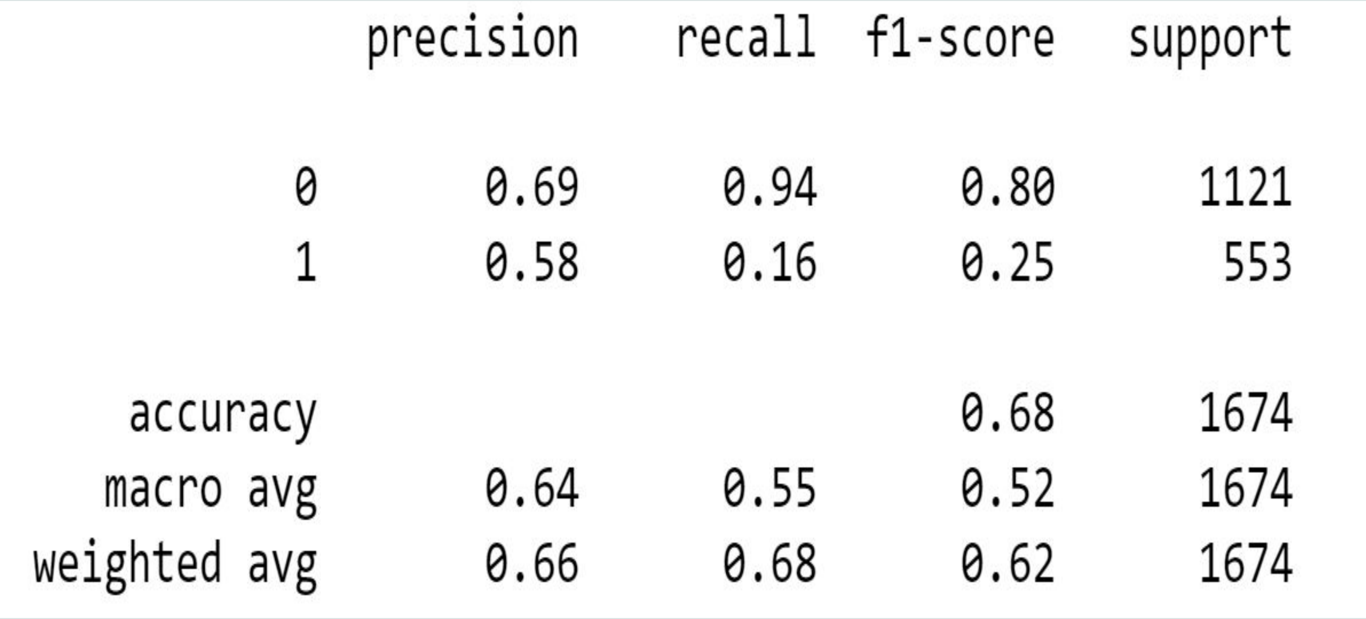


Data Modelling

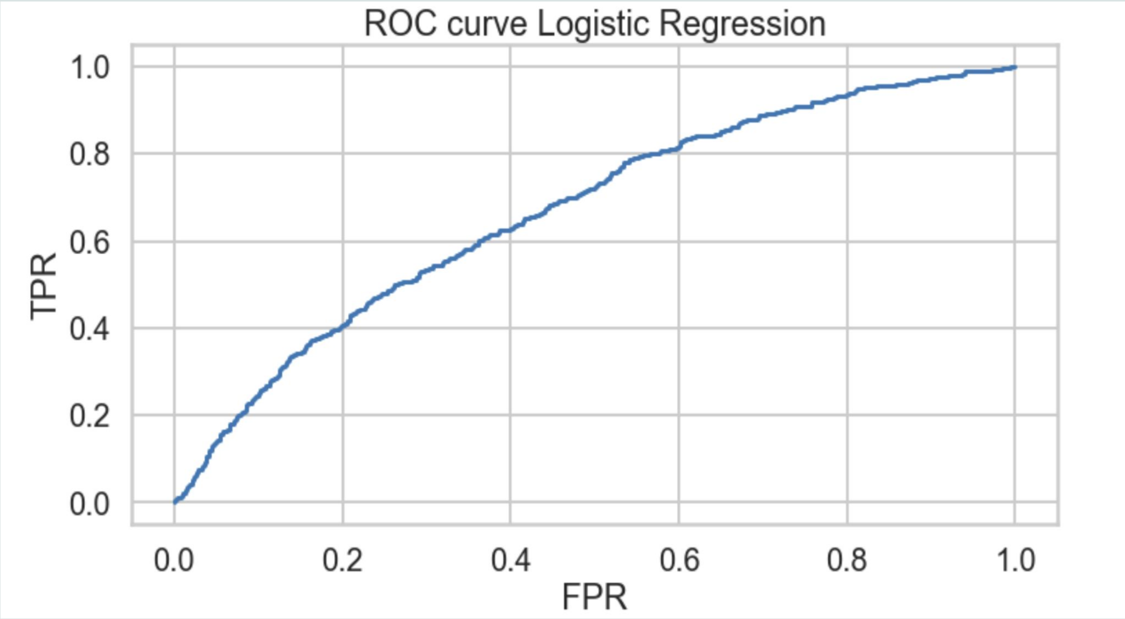


Logistic Regression

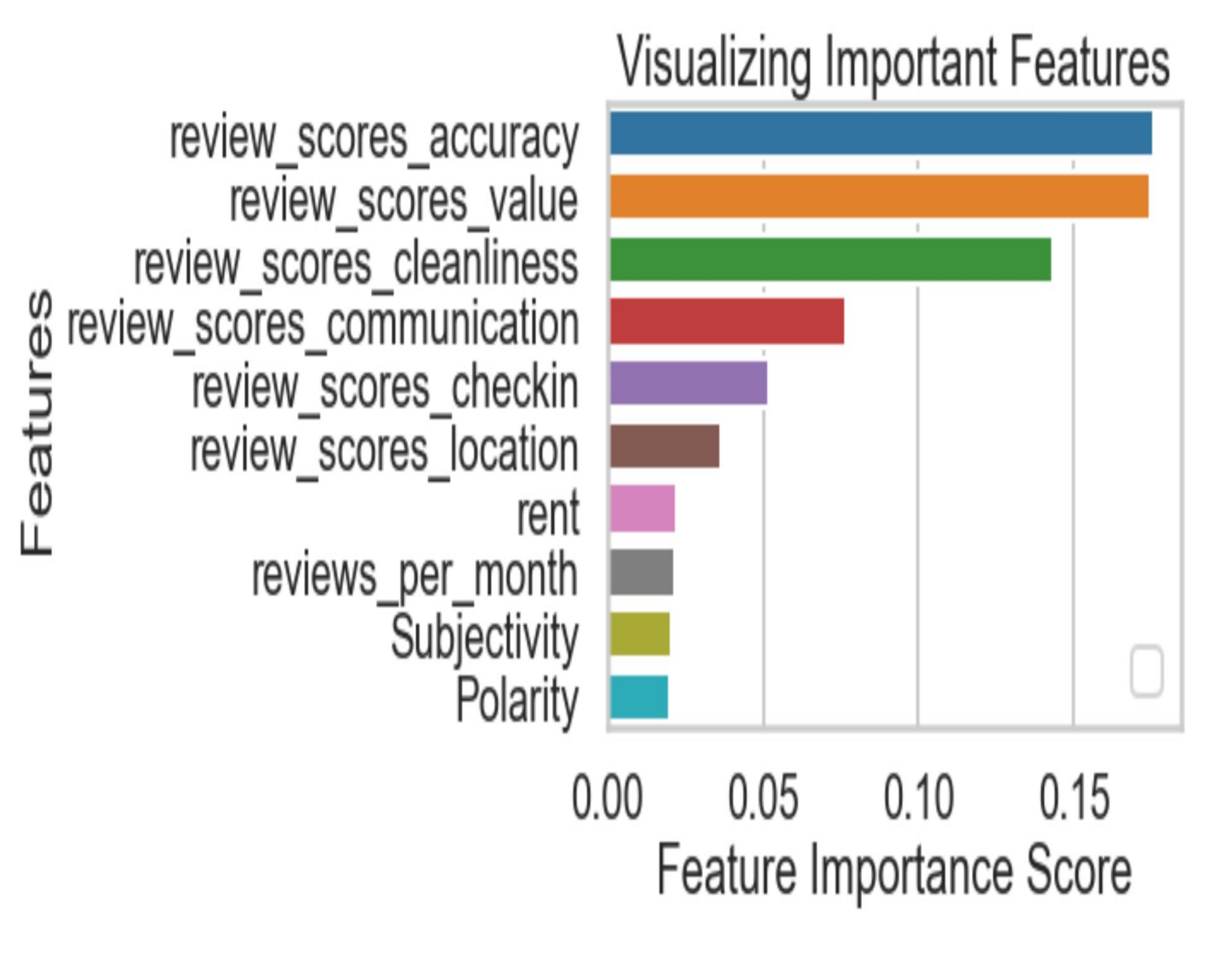
- Without processing the data
- Benchmark model



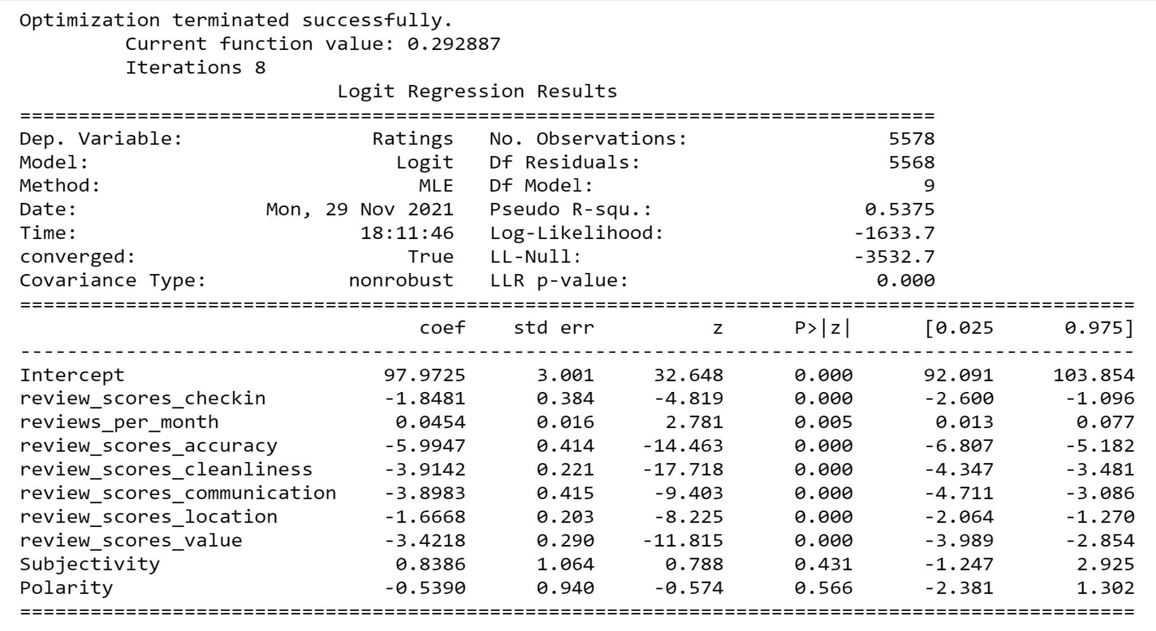
```
              precision    recall  f1-score   support

           0       0.69      0.94      0.80      1121
           1       0.58      0.16      0.25       553

    accuracy                           0.68      1674
   macro avg       0.64      0.55      0.52      1674
weighted avg       0.66      0.68      0.62      1674
```



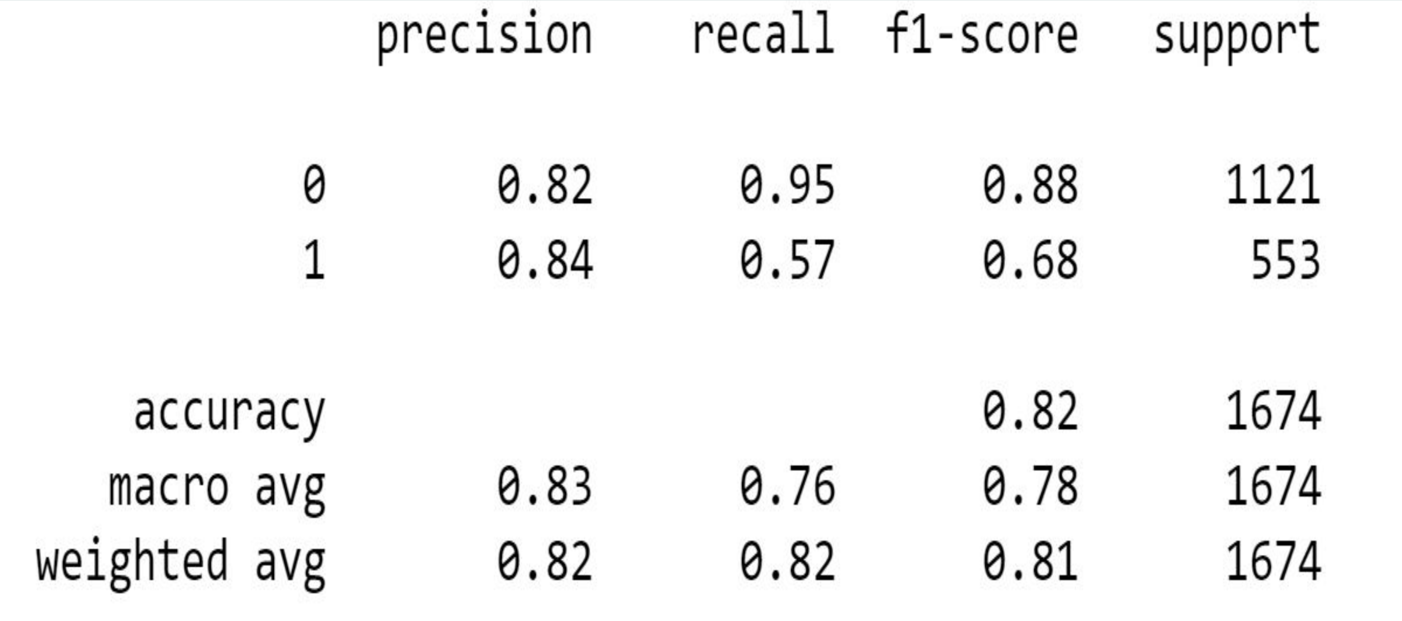
Feature Selection



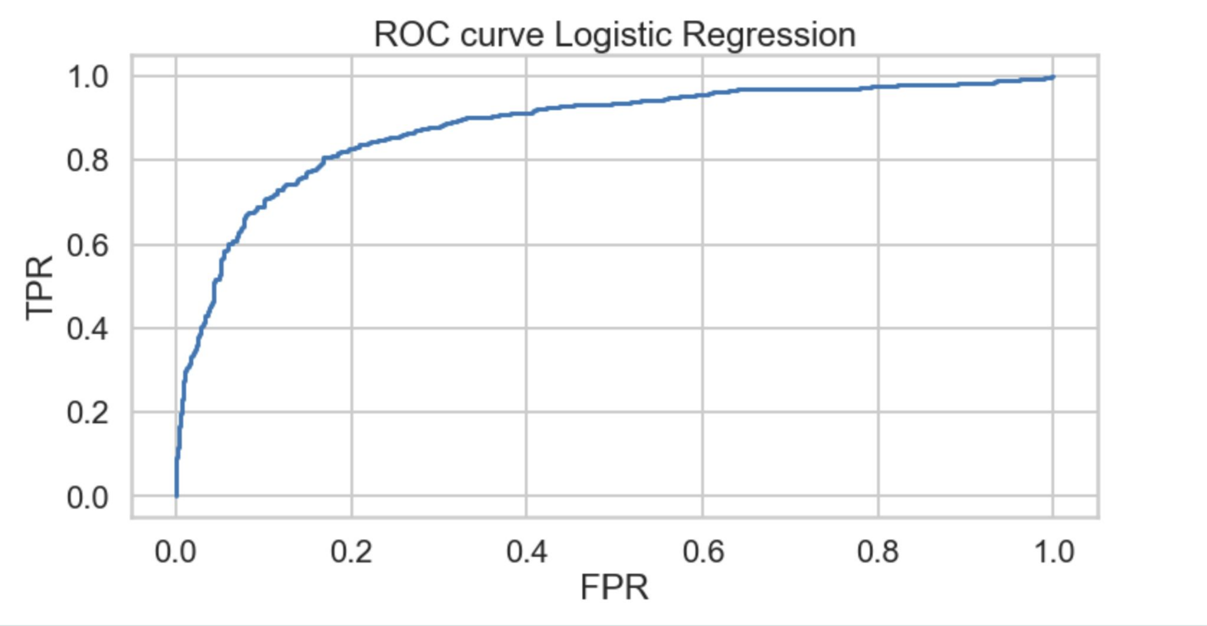
Logistic Regression with processed data



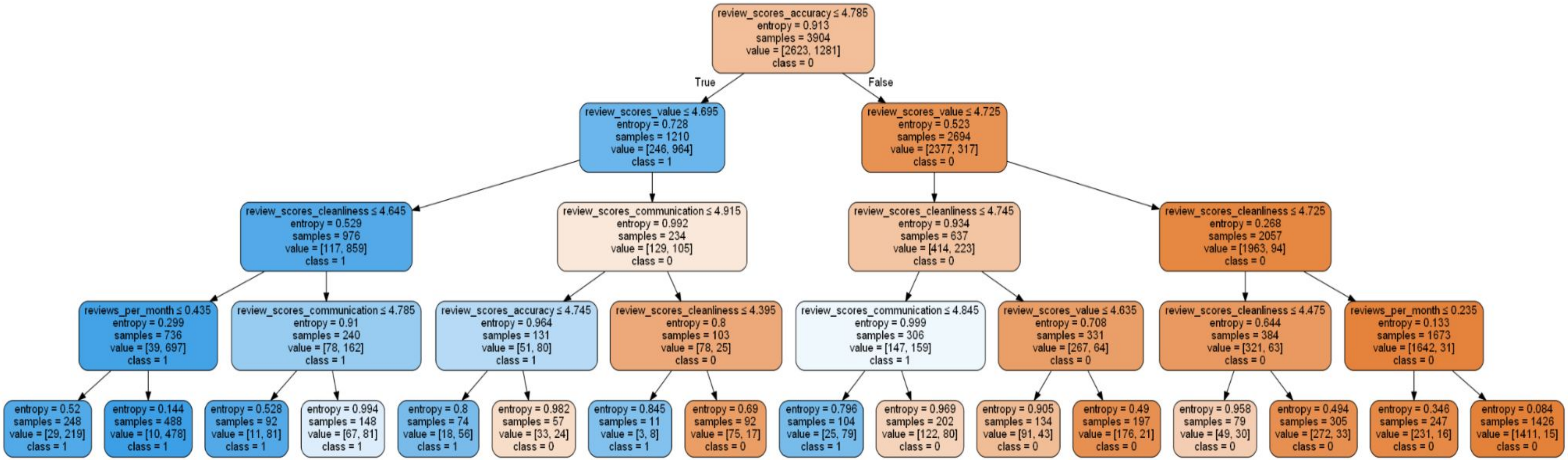
```
Optimization terminated successfully.
         Current function value: 0.292887
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:                Ratings   No. Observations:                 5578
Model:                          Logit   Df Residuals:                     5568
Method:                           MLE   Df Model:                            9
Date:                Mon, 29 Nov 2021   Pseudo R-squ.:                  0.5375
Time:                        18:11:46   Log-Likelihood:                -1633.7
converged:                       True   LL-Null:                       -3532.7
Covariance Type:            nonrobust   LLR p-value:                     0.000
===============================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                      97.9725      3.001     32.648      0.000      92.091     103.854
review_scores_checkin          -1.8481      0.384     -4.819      0.000      -2.600      -1.096
reviews_per_month               0.0454      0.016      2.781      0.005       0.013       0.077
review_scores_accuracy         -5.9947      0.414    -14.463      0.000      -6.807      -5.182
review_scores_cleanliness      -3.9142      0.221    -17.718      0.000      -4.347      -3.481
review_scores_communication    -3.8983      0.415     -9.403      0.000      -4.711      -3.086
review_scores_location         -1.6668      0.203     -8.225      0.000      -2.064      -1.270
review_scores_value            -3.4218      0.290    -11.815      0.000      -3.989      -2.854
Subjectivity                    0.8386      1.064      0.788      0.431      -1.247       2.925
Polarity                       -0.5390      0.940     -0.574      0.566      -2.381       1.302
===============================================================================================
```



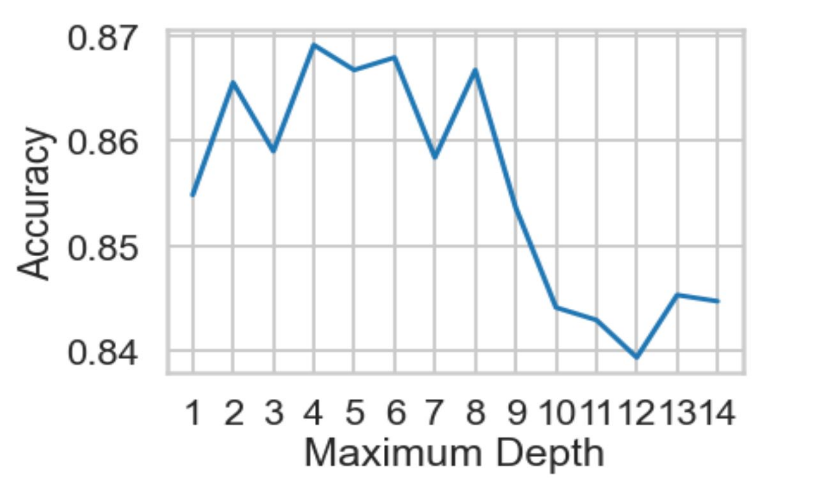
```
              precision    recall  f1-score   support

           0       0.82      0.95      0.88      1121
           1       0.84      0.57      0.68       553

    accuracy                           0.82      1674
   macro avg       0.83      0.76      0.78      1674
weighted avg       0.82      0.82      0.81      1674
```

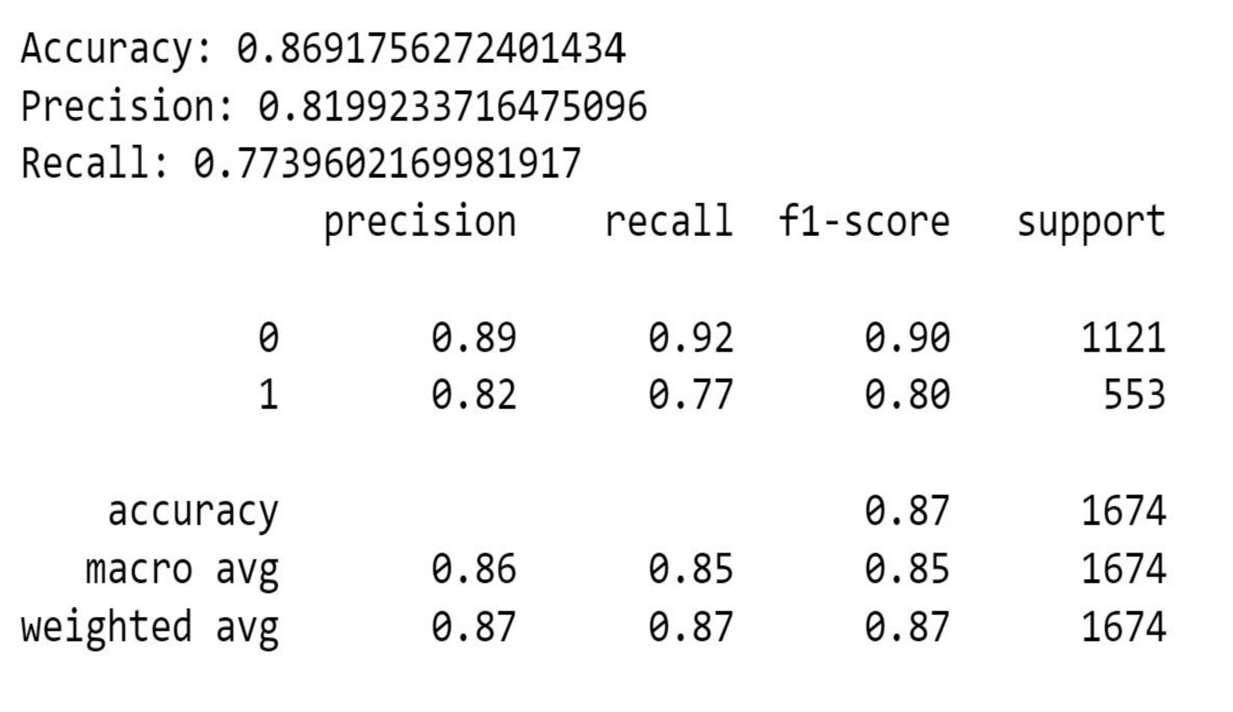


Decision Tree Classifier

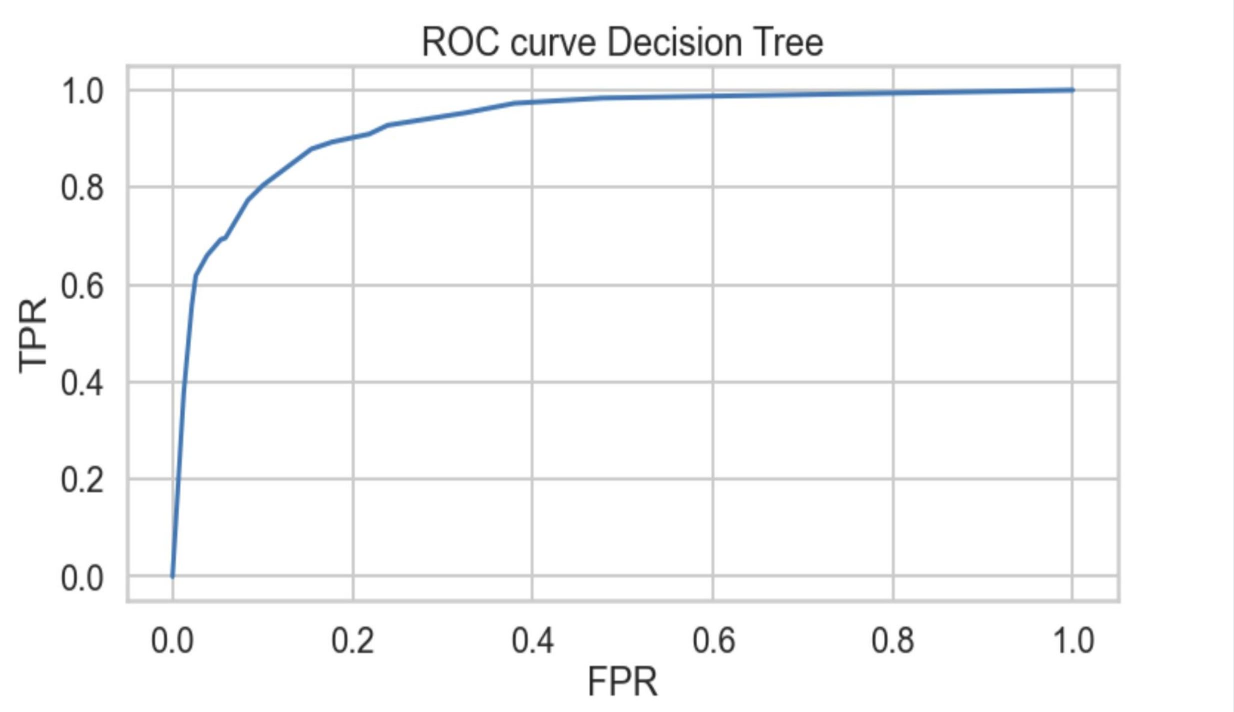


Decision Tree Classifier



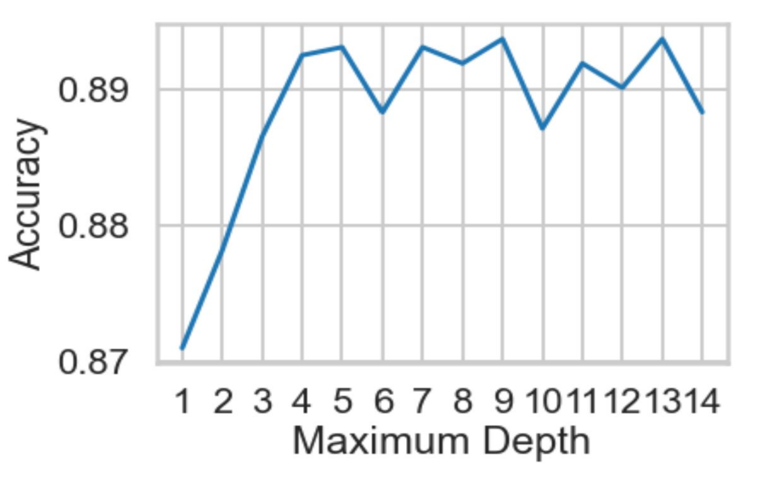


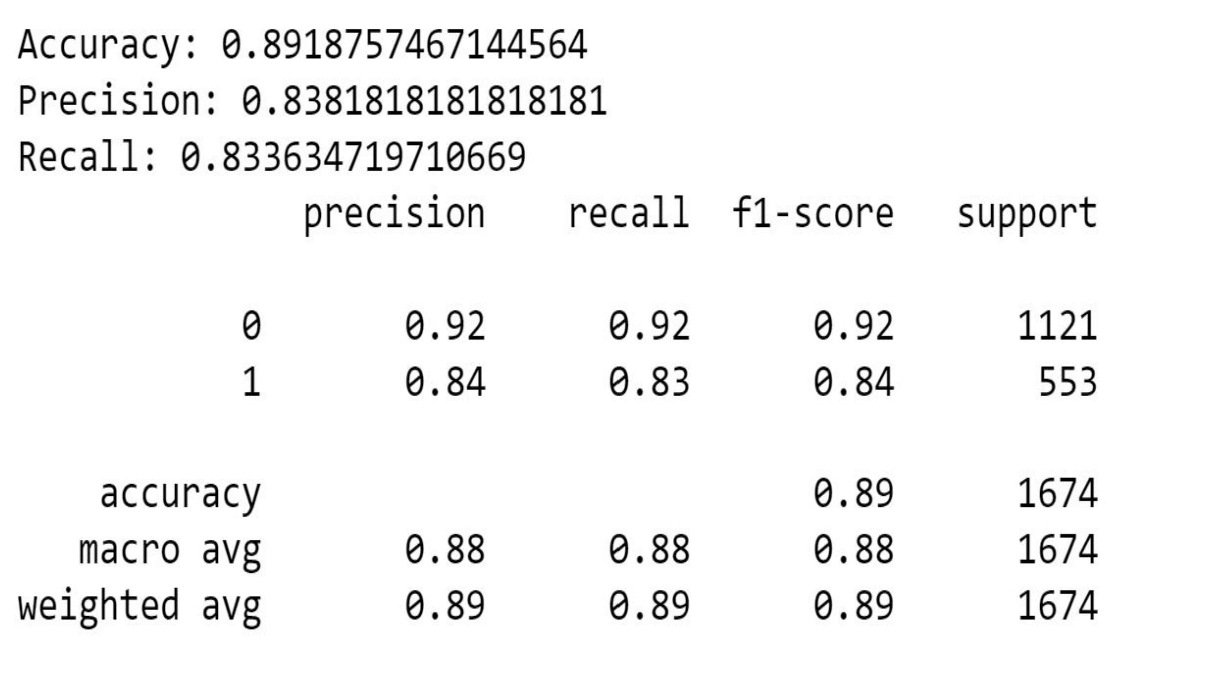
```
Accuracy: 0.8691756272401434
Precision: 0.8199233716475096
Recall: 0.7739602169981917
              precision    recall  f1-score   support

           0       0.89      0.92      0.90      1121
           1       0.82      0.77      0.80       553

    accuracy                           0.87      1674
   macro avg       0.86      0.85      0.85      1674
weighted avg       0.87      0.87      0.87      1674
```



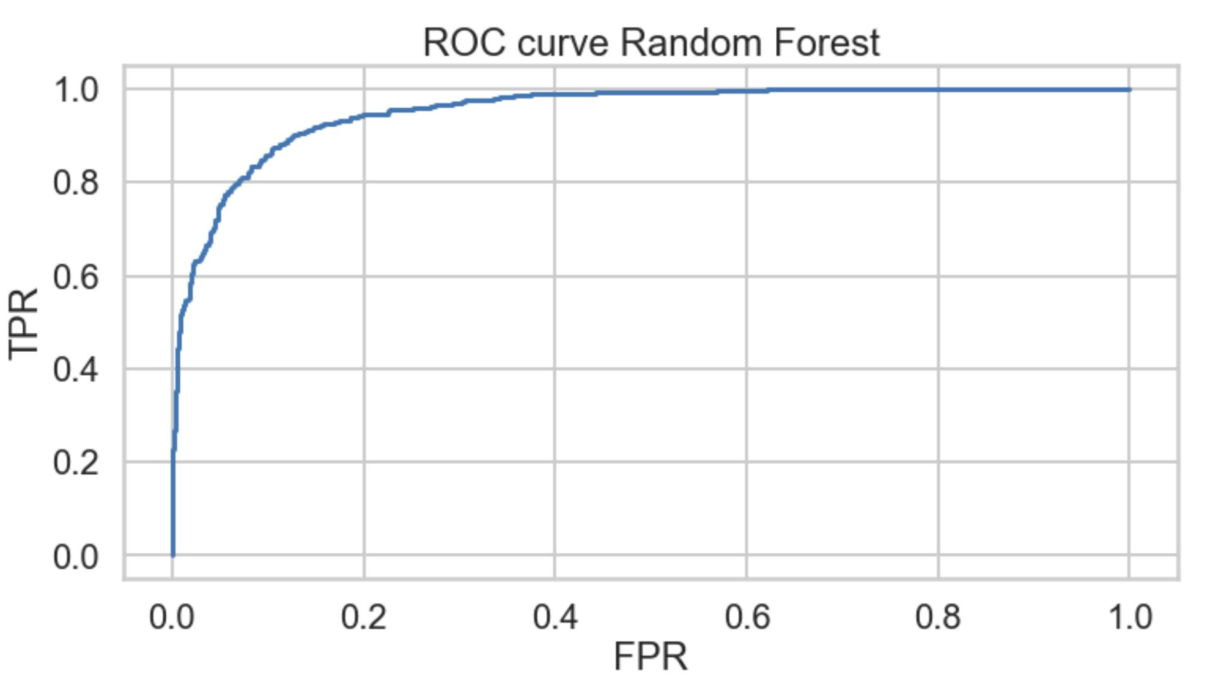
AUC - 0.9316823489747754

Random Forest



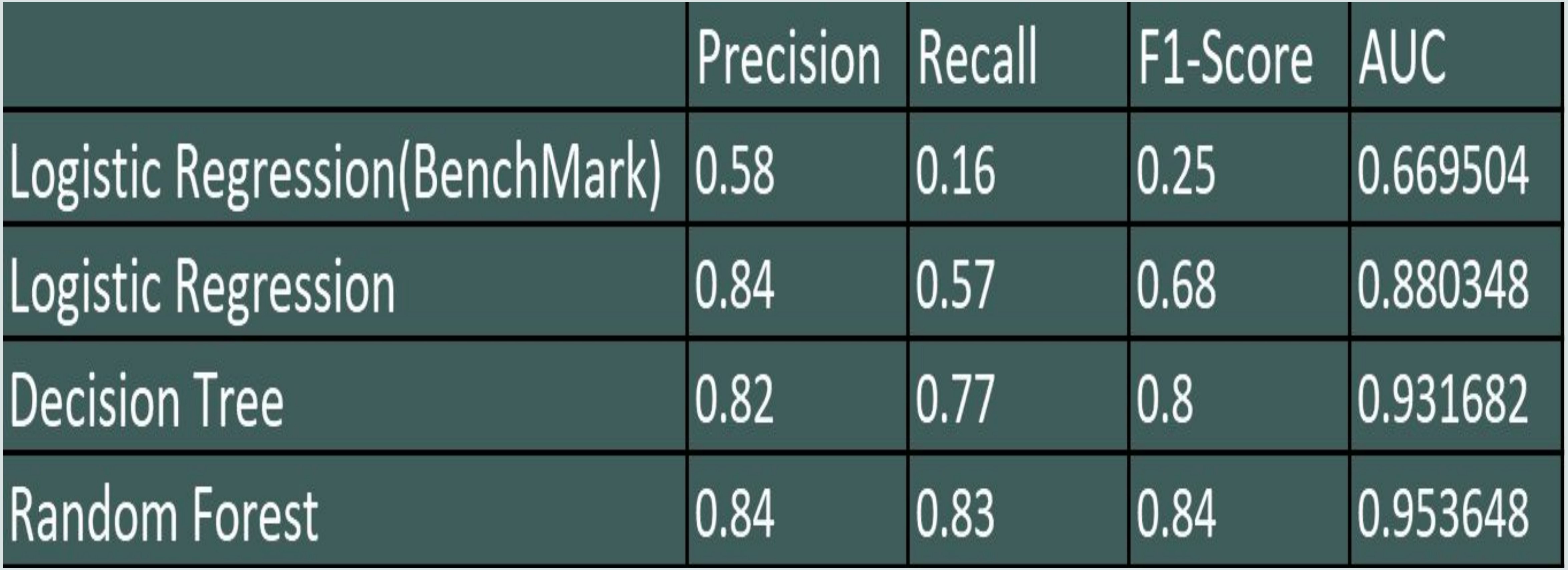


```
Accuracy: 0.8918757467144564
Precision: 0.8381818181818181
Recall: 0.833634719710669
              precision    recall  f1-score   support

           0       0.92      0.92      0.92      1121
           1       0.84      0.83      0.84       553

    accuracy                           0.89      1674
   macro avg       0.88      0.88      0.88      1674
weighted avg       0.89      0.89      0.89      1674
```



AUC - 0.9536483345243607

Conclusion



	Precision	Recall	F1-Score	AUC
Logistic Regression(BenchMark)	0.58	0.16	0.25	0.669504
Logistic Regression	0.84	0.57	0.68	0.880348
Decision Tree	0.82	0.77	0.8	0.931682
Random Forest	0.84	0.83	0.84	0.953648

KEY TAKEAWAYS

- Changing hyperparameters can change the accuracy of your model

- Avoid overfitting by limiting number of features/attributes used in prediction models

- We must make a decision regarding what attributes to use in model

  Sentiment analysis did not rate highly in significance but was integral to what we set out to do

Thank
You
Obrigado
Merci
Asante
Multumesc
Spasibo
Maake
Grazie
Kiitos
Kia Ora
Ua Tsaug Rau Koj
Matur Nuwun
Salamat
Mochchakkeram
Welalin
Matondo
Chokrane
Dankon
Mamana
Juspaxar
Nirringrazzjak
Raibh Maith Agat
Dank Je
Arigato
Vinaka
Cam on ban
Terma Kasih