Capstone Project

Airline Referral Prediction

By Nayana Pradeep

PROBLEM STATEMENT

- Given is a dataset of the Airline review details which has all the ratings for various services, route, aircraft name etc.

- Perform exploratory data analysis and visualization

- Create and experiment on various models to predict whether the reviewer refers the airline services to someone else.

OVERVIEW

Air transport enables millions of people to connect in cultural exchange.

It also boosts the tourism industry, which is a major economic factor both in the original countries and in the tourist destination countries.

Choosing the right airlines for a pleasant journey is crucial. Online reviews and friends' recommendation can influence this decision.

DATA SUMMARY

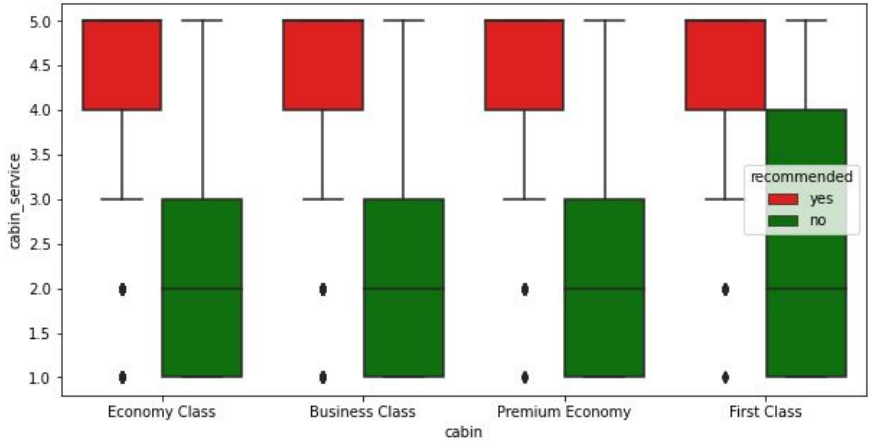
- Airline
- Overall – Ratings range is 1- 10
- Author
- Review Date
- Customer Review
- Aircraft
- Traveller type – Solo, Couple, Family and Business
- Cabin – Economy Class, Business class, Premium Economy and First Class
- Date Flown
- Seat Comfort – Ratings range is 1- 5
- Cabin Service – Ratings range is 1- 5
- Food and Beverage – Ratings range is 1- 5
- Entertainment – Ratings range is 1- 5
- Ground Service – Ratings range is 1- 5
- Value for money – Ratings range is 1- 5
- Recommended – Yes or No

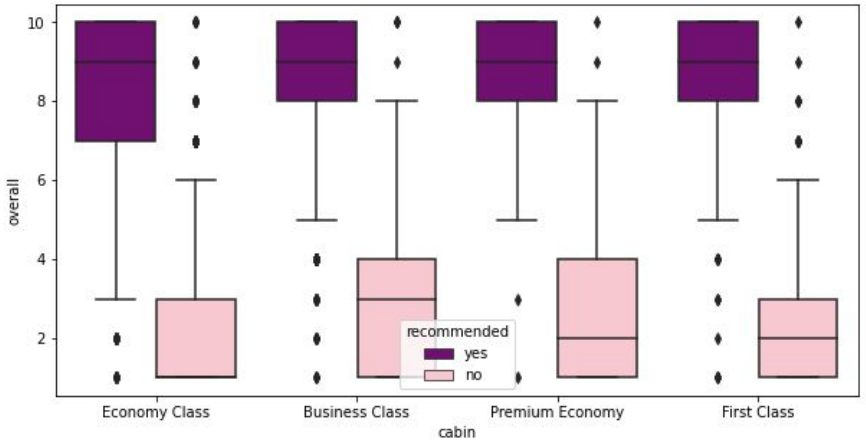
NULL VALUES TREATMENT

- 131895 rows and 17 columns
- Huge amount of null values
- Dropped all the rows with recommended as null values
- Dropped the columns ‘author’, ‘time_flown’ and ‘aircraft’
- Dropped all the rows with null values
- Dropped all the duplicate rows
- Dataset reduced to 17 percent of the original dataset. But the data is clean

EDA AND VISUALIZATION

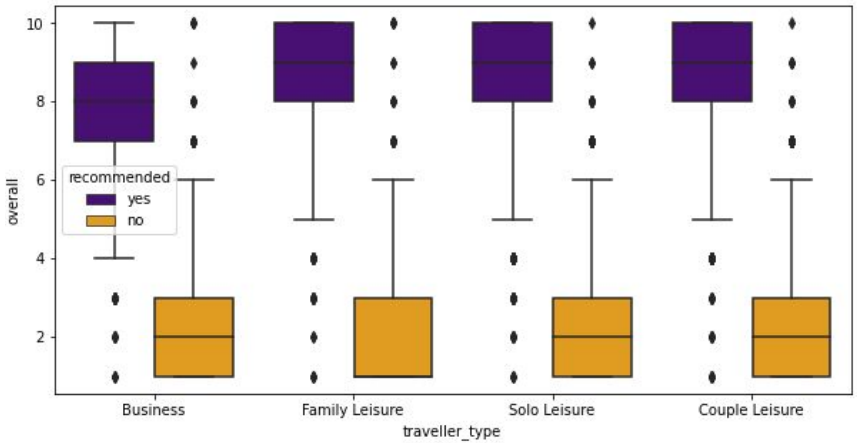
- CABIN TYPE

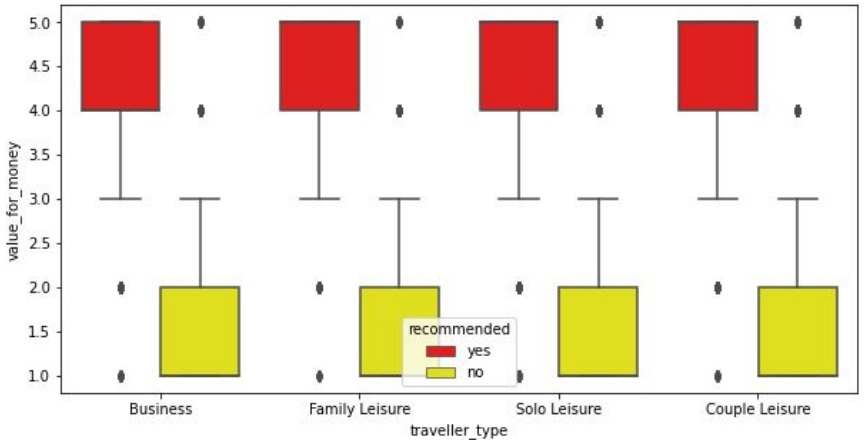




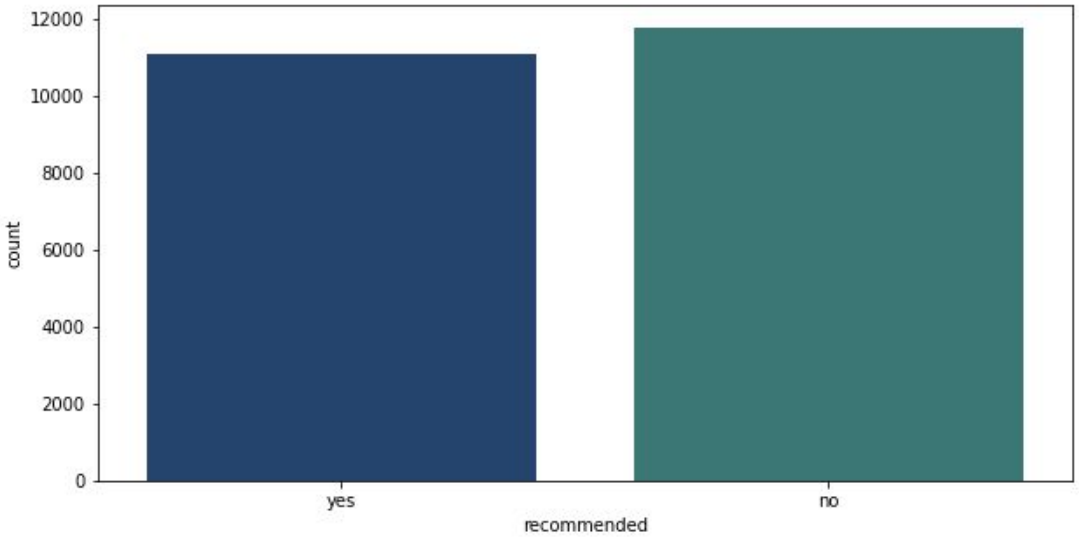
EDA AND VISUALIZATION

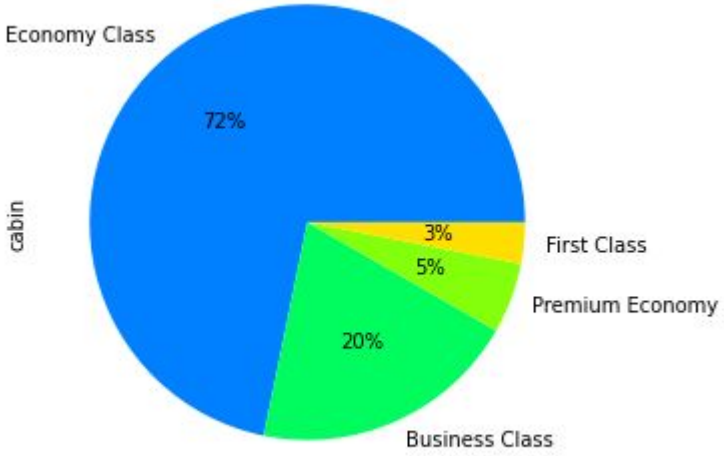
- TRAVELLER TYPE



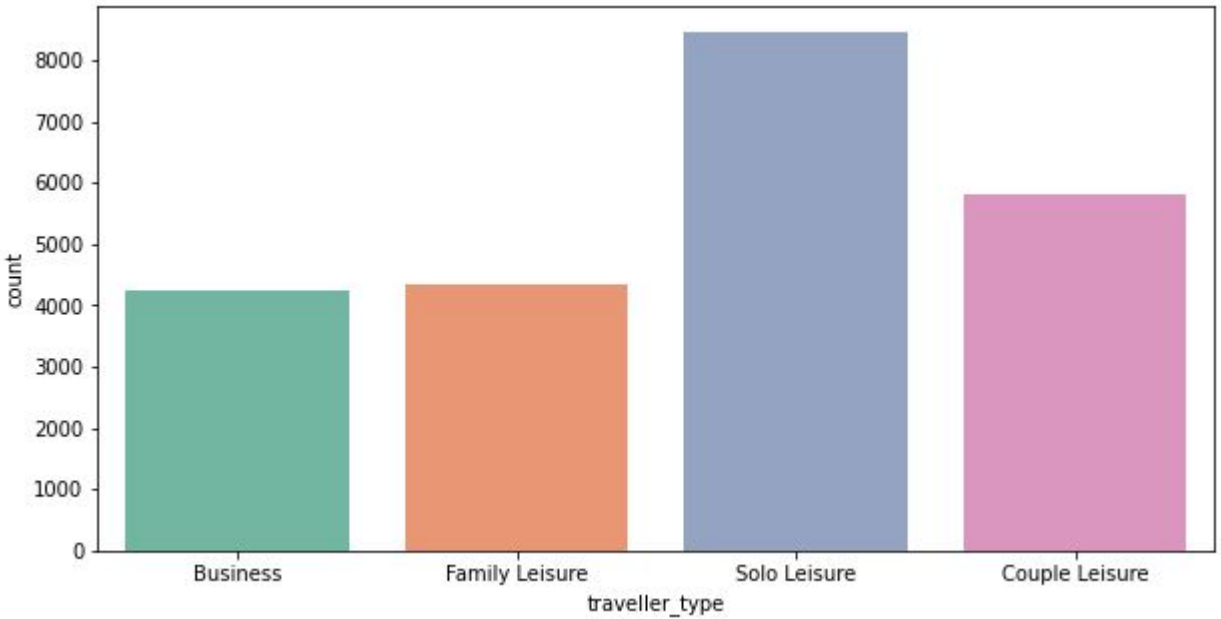


EDA AND VISUALIZATION



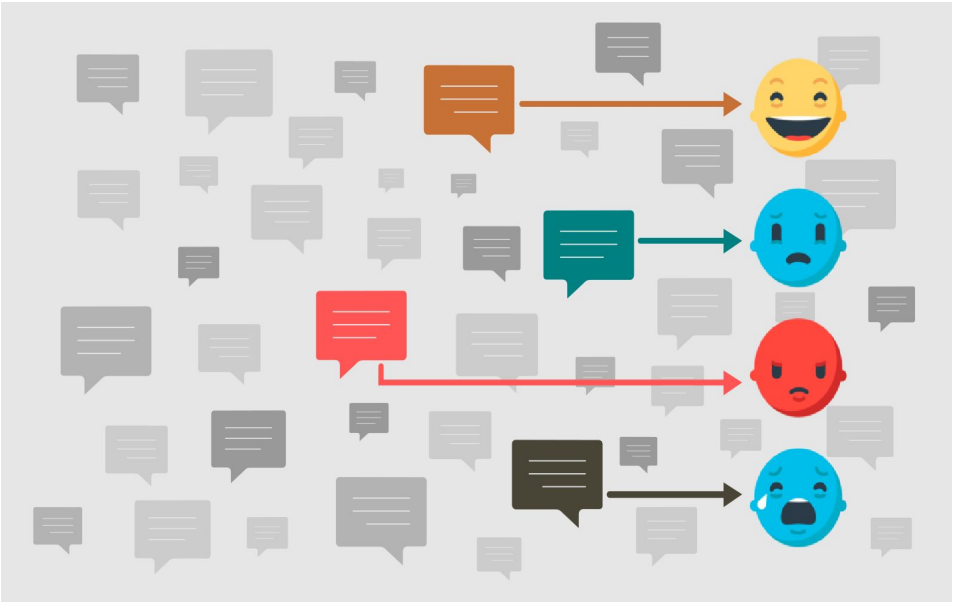


EDA AND VISUALIZATION



NLP –SENTIMENT ANALYSIS

- Performed on customer review column
- Determines whether a piece of writing is positive, negative or neutral.
- Combines NLP and machine learning techniques
- Assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase.



SENTIMENT ANALYSIS –TEXT CLEANING

- Convert the text to lower case
- Tokenize the text
- Remove useless words that contain numbers
- Remove useless stop words like ‘the’, ‘a’ ,’this’ etc.
- Part-Of-Speech (POS) tagging
- Lemmatize the text

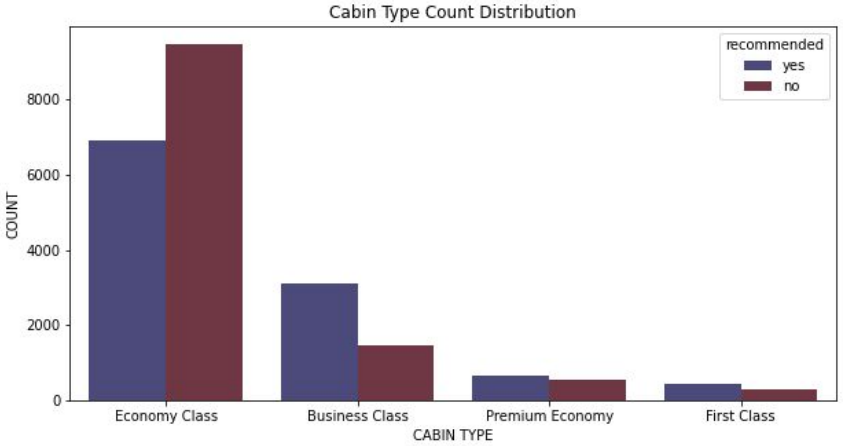
SENTIMENT ANALYSIS – SCORES

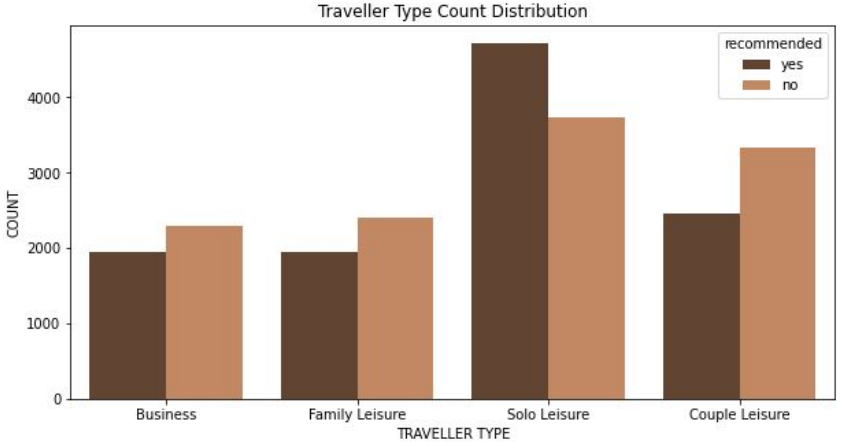
Vader is a part of the NLTK module designed for sentiment analysis. Vader uses a lexicon of words to find which ones are positives or negatives. It also takes into account the context of the sentences to determine the sentiment scores.

Following scores are determined from the cleaned text :

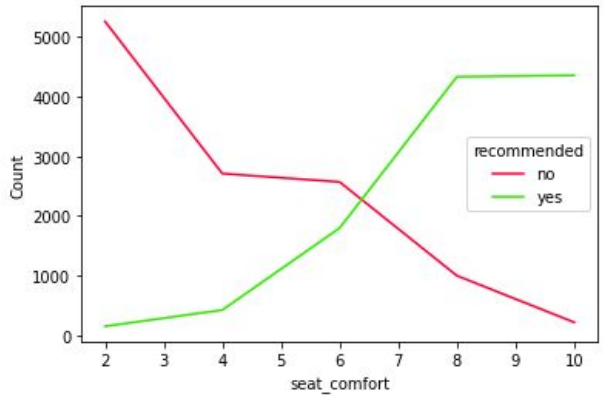
- a neutrality score
- a positivity score
- a negativity score
- an overall score that summarizes the previous scores

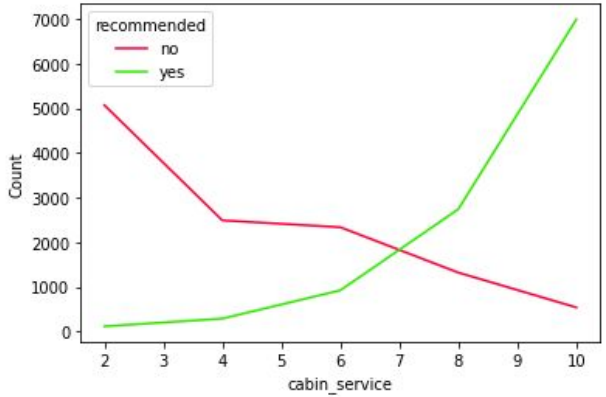
RECOMMENDATION STATUS

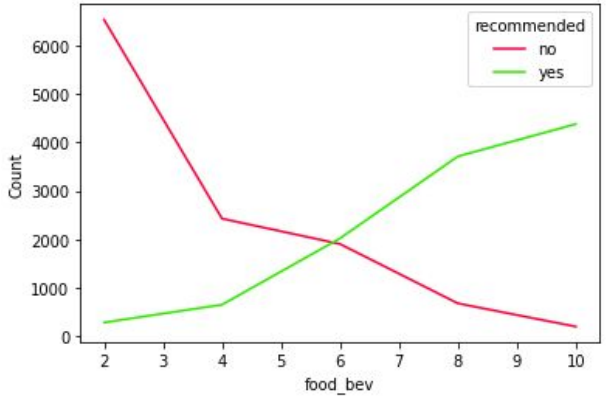


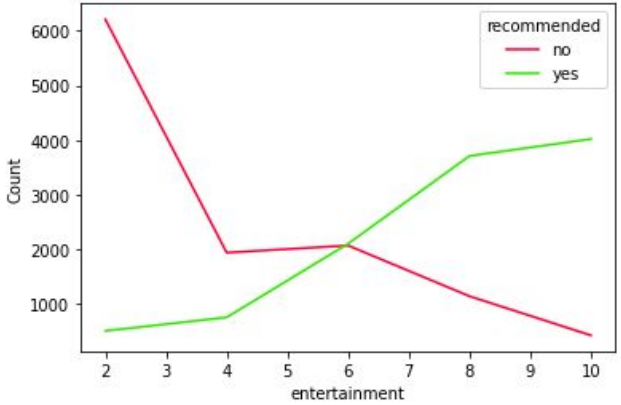


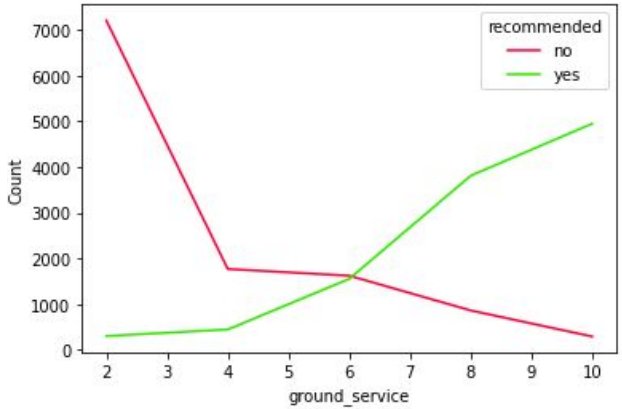
RECOMMENDATION AND RATINGS

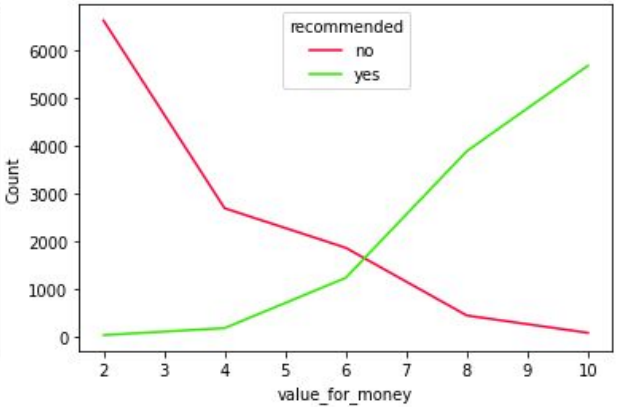






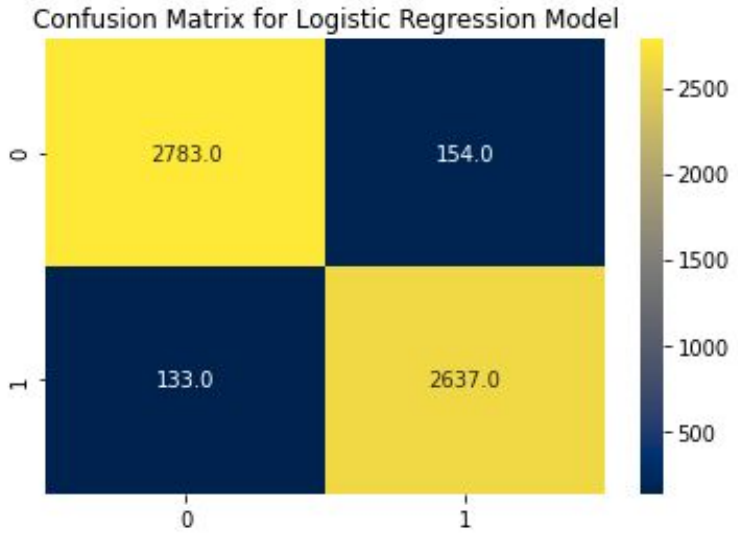






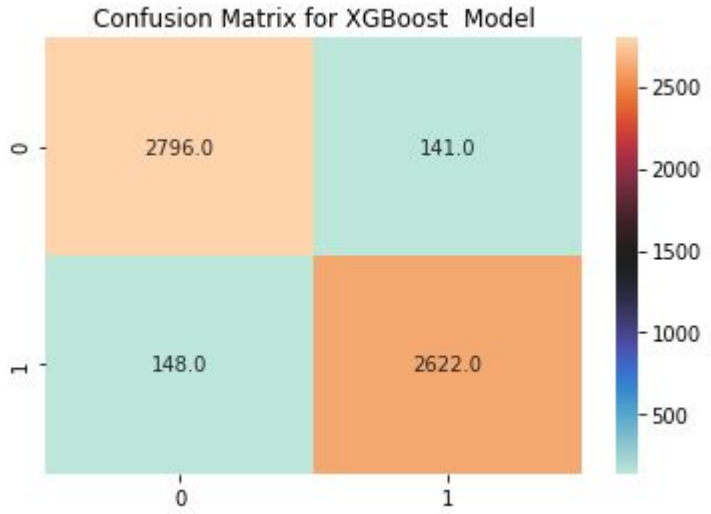
LOGISTIC REGRESSION MODEL

- Accuracy : 0.949711
- Recall : 0.951986
- Precision : 0.944823
- F-1 Score : 0.948391
- ROC AUC Score : 0.949776



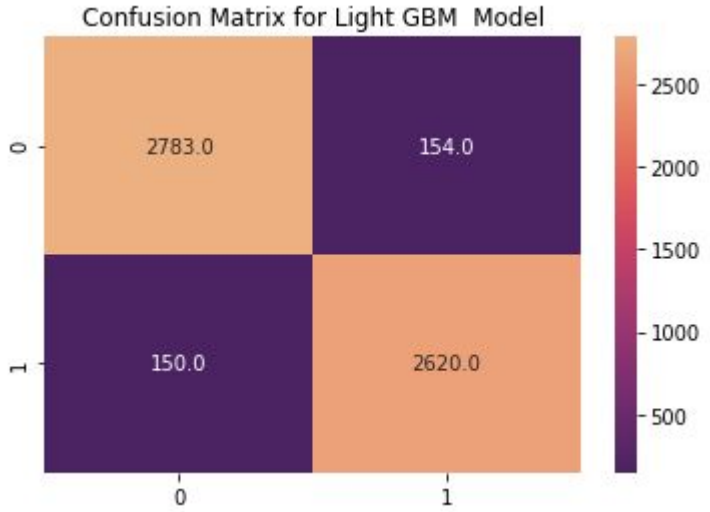
XGBOOST MODEL

- Accuracy : 0.949360
- Recall : 0.946570
- Precision : 0.948969
- F-1 Score : 0.947768
- ROC AUC Score : 0.949281



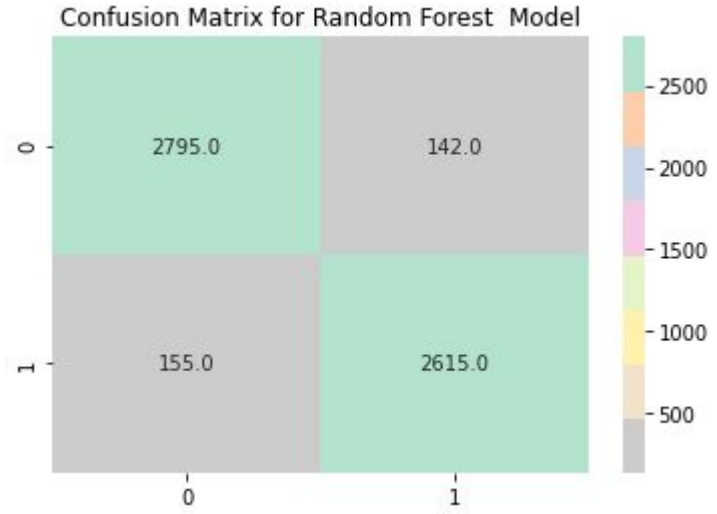
LIGHT GBM MODEL

- Accuracy : 0.946732
- Recall : 0.945848
- Precision : 0.944484
- F-1 Score : 0.945166
- ROC AUC Score : 0.946707



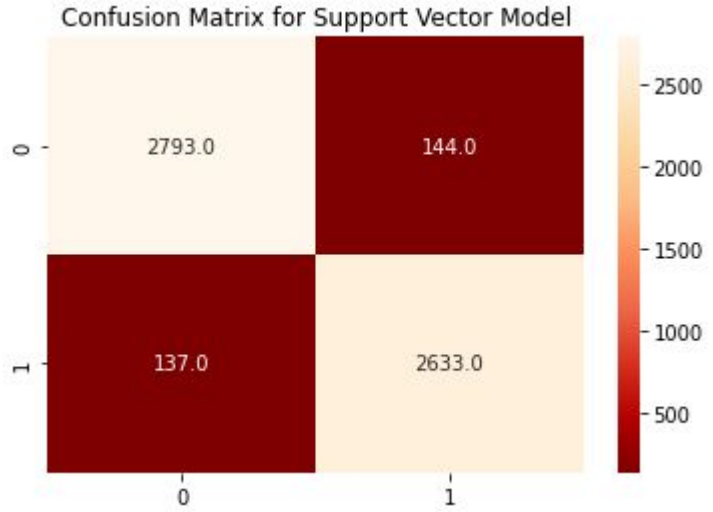
RANDOM FOREST MODEL

- Accuracy : 0.948134
- Recall : 0.942960
- Precision : 0.949818
- F-1 Score : 0.946377
- ROC AUC Score : 0.947987



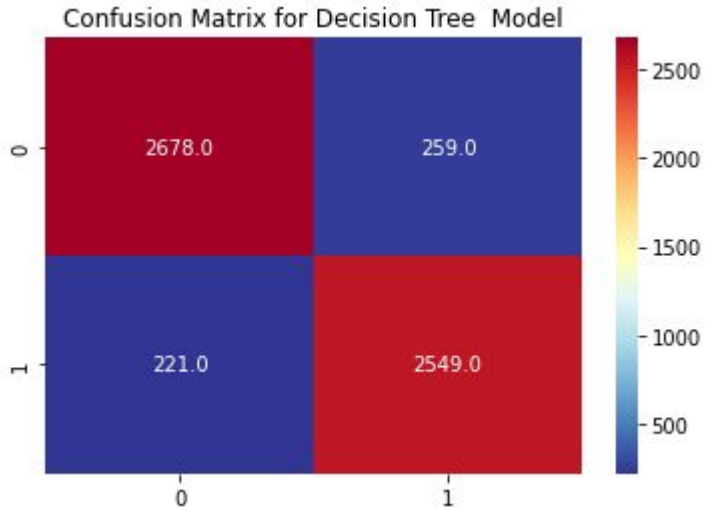
SUPPORT VECTOR MACHINE

- Accuracy : 0.950762
- Recall : 0.950542
- Precision : 0.948145
- F-1 Score : 0.949342
- ROC AUC Score : 0.950756



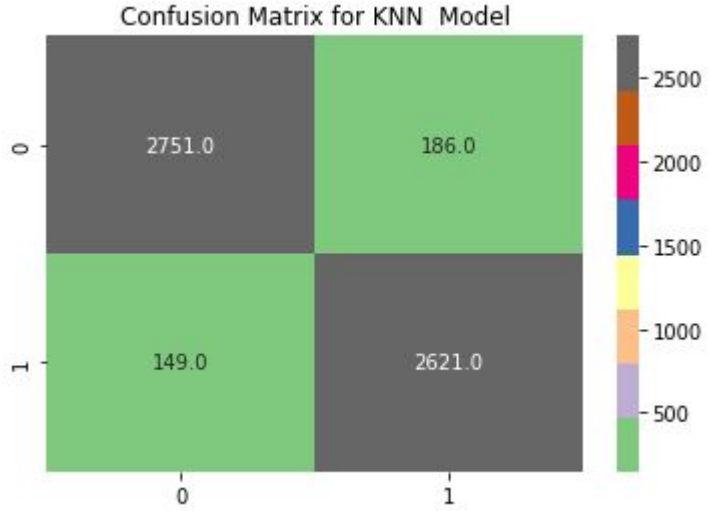
DECISION TREE MODEL

- Accuracy : 0.915893
- Recall : 0.920217
- Precision : 0.907764
- F-1 Score : 0.913948
- ROC AUC Score : 0.916016



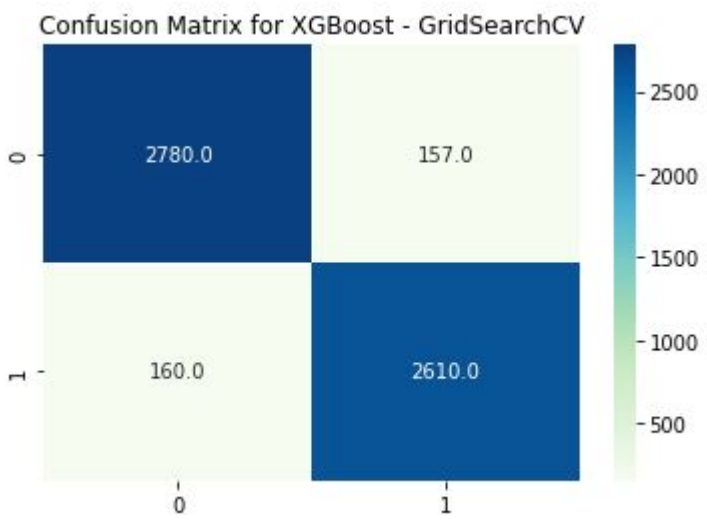
K-NEAREST NEIGHBOUR MODEL (KNN)

- Accuracy : 0.941300
- Recall : 0.946209
- Precision : 0.933737
- F-1 Score : 0.939932
- ROC AUC Score : 0.941440



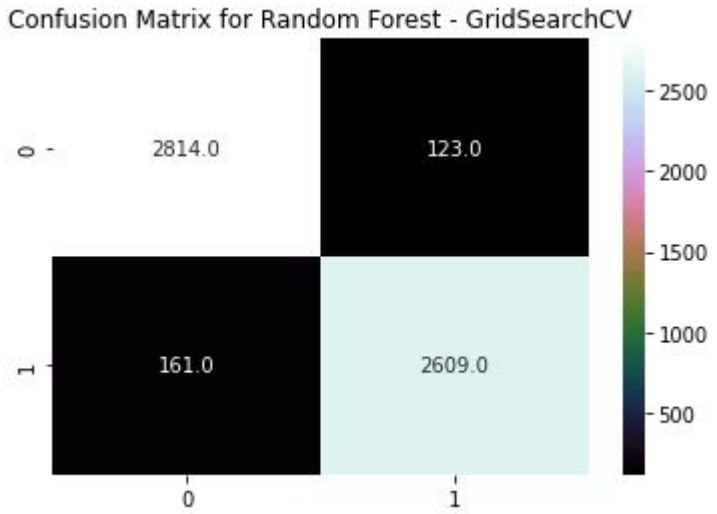
XGBOOST - GRIDSEARCHCV

- Accuracy : 0.944454
- Recall : 0.942238
- Precision : 0.943260
- F-1 Score : 0.942749
- ROC AUC Score : 0.944391



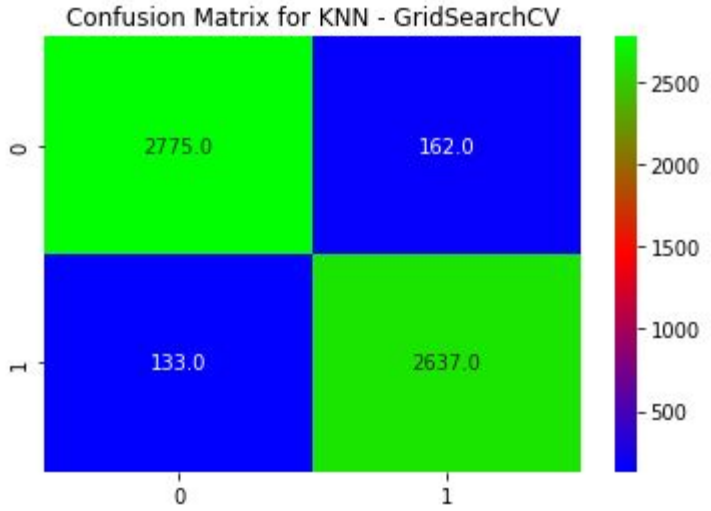
RANDOM FOREST - GRIDSEARCHCV

- Accuracy : 0.950237
- Recall : 0.941877
- Precision : 0.954978
- F-1 Score : 0.948382
- ROC AUC Score : 0.949999



KNN - GRIDSEARCHCV

- Accuracy : 0.948309
- Recall : 0.951986
- Precision : 0.942122
- F-1 Score : 0.947028
- ROC AUC Score : 0.948414



MODEL COMPARISON

	MODEL NAME	ACCURACY	RECALL	PRECISION	F1-SCORE	ROC AUC SCORE
0	Support Vector Machine	95.08	95.05	94.81	94.93	95.08
1	Random Forest - GridSearchCV	95.02	94.19	95.50	94.84	95.00
2	Logistic Regression	94.97	95.20	94.48	94.84	94.98
3	XGBoost	94.94	94.66	94.90	94.78	94.93
4	KNN - GridSearchCV	94.83	95.20	94.21	94.70	94.84
5	Random Forest	94.81	94.30	94.98	94.64	94.80
6	Light GBM	94.67	94.58	94.45	94.52	94.67
7	XGBoost - GridSearchCV	94.45	94.22	94.33	94.27	94.44
8	KNN Model	94.13	94.62	93.37	93.99	94.14
9	Decision Tree	91.59	92.02	90.78	91.39	91.60

CONCLUSION

- **Support Vector Machine** has the best accuracy among the experimented models even thought it was a very close call.

- All experimented models were having almost same accuracy rate above 94% except for the decision tree.

CHALLENGES

- Large dataset with huge amount of null values
- Analyzing how to clean the data without losing the significant data
- Close proximity of the evaluation scores of the experimented models

THANK YOU