Customer Churn Prediction In A Telecommunication Company Using Machine Learning Algorithms.

***Abstract*— Customer retention is essential to the success and growth of any business. A study conducted by Bain & Company and Earl Sasser revealed that a 5% increase in customer retention improves profitability by 25% [5]. Another study showed that it costs five times as much to attract new customers than to retain existing ones [1]. As a result, organizations are looking to develop algorithms that can predict which customers are more likely to leave (churn) and take appropriate action. This study analyzes data from a telecommunication company with the aim of predicting customer churn using machine learning algorithms and proposing relevant retention strategies. Six machine learning classification algorithms were implemented: Logistic Regression, Decision Tree, Random Forest, Support Vector Classifier, Light Gradient Boosting Machine Classifier and Artificial Neural Network. The result shows that Random Forest outperforms other models.**

I. Introduction

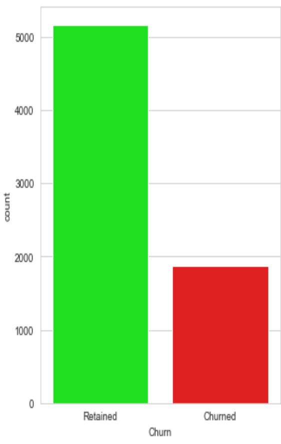
Several decades ago, before the advent of the industrial revolution and the coming of age of the internet, man's basic needs were simply food, shelter and clothing. With the world transitioning into a much smaller entity as the years roll by, internet and phone services are fast becoming basic needs to live a quality life. The global telecom market is expected to reach $3461.03 billion in 2025 [6]. With increased players in the telecommunications industry, there is a fiercer competition for customers. While losing a single customer might not hurt a business, not identifying a pattern and accurately predicting which customers are likely to churn can lead to a sizeable loss in revenue. Zhang T. et al. [7] investigated customer churn in three Chinese telecom companies and found that logistic regression performed better than Fisher's discriminant. Zhang X. et al [9] researched on customer churn based on customer segmentation, using SAS data mining technology and fuzzy C-means clustering algorithm. Mahajan et al. [4] discovered that Prepaid customer attrition is heavily influenced by prices and service quality, according to a factor analysis and regression analysis. As a result, businesses must assess and enhance their service quality.

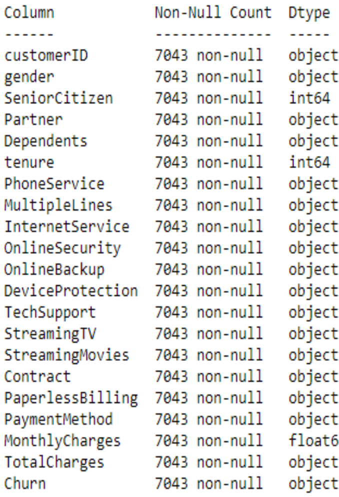
II. Dataset Exploration and features selection

A. Dataset Description

The Data from this study is obtained from kaggle. The dataset provides information about the customers in a telecommunications company. Each row represents a customer, and the columns contain customer attributes, resulting in 7043 rows and 21 columns. The dataset includes information about:

- Customers who have left in the previous (churn column)
- Services used by the customer.
- Account information of customers.
- Information about clients' demographics.





```
Column            Non-Null Count  Dtype
------            --------------  -----
customerID        7043 non-null   object
gender            7043 non-null   object
SeniorCitizen     7043 non-null   int64
Partner           7043 non-null   object
Dependents        7043 non-null   object
tenure            7043 non-null   int64
PhoneService      7043 non-null   object
MultipleLines     7043 non-null   object
InternetService   7043 non-null   object
OnlineSecurity    7043 non-null   object
OnlineBackup      7043 non-null   object
DeviceProtection  7043 non-null   object
TechSupport       7043 non-null   object
StreamingTV       7043 non-null   object
StreamingMovies   7043 non-null   object
Contract          7043 non-null   object
PaperlessBilling  7043 non-null   object
PaymentMethod     7043 non-null   object
MonthlyCharges    7043 non-null   float64
TotalCharges      7043 non-null   object
Churn             7043 non-null   object
```

Figure 1. (target column)   Figure 2. (data description)

B. Data Exploration and Recommendations

After exploratory data analysis, the Figure 1 shows that approximately 27% of the customers churned. The following points were highlighted from Exploratory data analysis:

- There are younger and middle-aged customers than senior citizens.
- Many customers are leaving in the first month. This is something the company needs to investigate.
- Customers without dependents are less likely to churn.
- Customers who use a phone service are less likely to churn.

C. Data Processing and Feature selection

There was no missing value observed in the data, although the 'TotalCharges' column had an object datatype and was converted to float. Analysis of the continuous variables using a correlation matrix show that the 'TotalCharges' column was highly correlated with other variables and so was dropped (appendix). Analysis of the categorical variables using their Chi-square and p-values was used to further reduce the features. The categorical variables in the dataset were then encoded using one-hot encoding to represent the data in a binary string. For the dependent Variable, the response "NO" was converted to 0 and the response "YES" to 1.

III. EXPERIMENTS

The aim of the study is predicting whether a customer would churn or not, i.e., classifying customers into two classes (binary classification). Following the data preprocessing, churned customers accounted for just 27% of the target variable. SMOTE ('Synthetic Minority Oversampling Technique') was used to deal with this imbalance. This approach aids in mitigating the effects of overfitting caused by random oversampling. The scaled data was then trained and tested with six supervised machine learning algorithms:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machines (SVC)
- Light Gradient Boosting Machine (LGBM)
- Artificial Neural Networks

IV. EVALUATION METRICS

The evaluation of models is an essential part of the modelling process. It helps us access which model is best for representing our data and determining the models capacity to handle unfamiliar data. The following evaluation metrics have been chosen:

A. Precision

Precision indicates the reliability of the model's positive prediction. It refers to the number of true positives divided by the total number of positive predictions. In this scenario, it divides the total number of consumers anticipated to churn by the number of customers properly forecasted to churn.

$$\text{Precision} = \frac{TP}{TP+FP}$$

B. Recall

Recall is the ratio between the true positives and the sum of true positives and false negatives. It attempts to answer the question: what percentage of true positives was successfully identified?

$$\text{Recall} = \frac{TP}{TP+FN}$$

C. F1 Score

F1 score is the harmonic mean of the model's precision and recall. It is represented by:

$$\text{F1} = 2\ x\ \frac{Precision\ x\ Recall}{precision+Recall}$$

D. Accuracy

Accuracy is the fraction of predictions our model got right. It describes how the model performs across all classes. It is represented by the ratio between the correct predictions and the total number of predictions

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

E. AUROC and ROC Curve

The Receiver Operating Characteristics curve is abbreviated as ROC and AUROC is the area under the curve. "The ROC curve summarizes the prediction performance of a classification model at all classification thresholds, as a function of the True positive Rate (TPR) and the False Positive Rate (FPR)" [11]. It shows how well the model is capable of distinguishing between the two classes.

V. RESULTS

The table below shows a summary of the results, based on the evaluation metrics.

Model	Class	PREC	REC	F1	ACC	AUC
Logistic Regression	Retained	0.81	0.79	0.80	0.81	0.901
	Churned	0.81	0.83	0.82	0.81	
Decision Trees	Retained	0.78	0.78	0.79	0.79	0.798
	Churned	0.79	0.79	0.80	0.79	
Random Forest	Retained	0.83	0.86	0.84	0.84	0.920
	Churned	0.86	0.84	0.85	0.84	
SVC	Retained	0.81	0.82	0.82	0.82	0.900
	Churned	0.83	0.82	0.82	0.82	
LGBM Classifier	Retained	0.84	0.83	0.83	0.84	0.919
	Churned	0.84	0.85	0.84	0.84	
ANN	Retained	0.81	0.81	0.81	0.81	
	Churned	0.82	0.81	0.82	0.81	

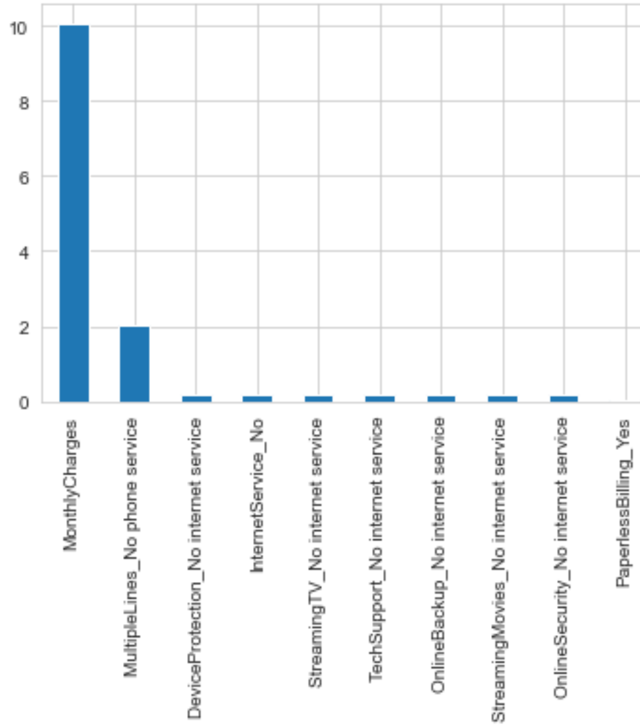


Figure 3. Weights of dependent to independent variable

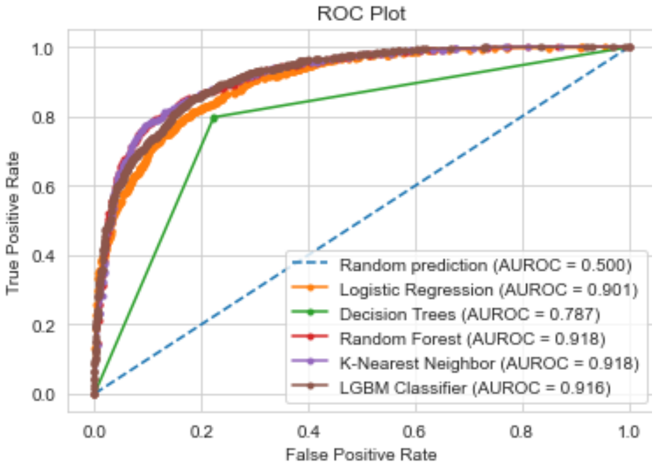


Figure 4. ROC Curve.

An increase in monthly charges will most likely result in a customer churning. To retain customers, the company can consider granting new subscribers a few months free access to the service. Random Forest was the best performing algorithm with an accuracy of 84% and AUROC of 0.918. Attempt was made at improving the model by tuning its hyperparameters using Grid Search, but it produced similar results.

VI. Discussion and Justification

A. Random Forest

Random Forest is an ensemble of decision trees, trained via the bagging method. When building trees, it adds extra randomization by looking for the best feature from a random subset of features [3]. Tin kam Ho [8] demonstrated that the problem of generalization can be overcome using multiple trees. He showed that random Forest increases classifier complexity without trading off generalization accuracy. It is no surprise that it performed better than Decision trees classifier in this model. Random Forest also makes it easier to measure the relative importance of each feature by calculating the weighted average.

B. LGBM Classifier

LightGBM is a decision tree-based gradient boosting framework that is fast, distributed, and high-performing. Gradient boosting works by sequentially adding predictors to an ensemble. LightGBM is capable of handling large-scale data, uses lower memory and achieves high efficiency. Guolin Ke et al., showed that LGBM outperforms XGBoost in terms of computational speed and memory consumption, while maintaining accuracy [10].

C. Artificial Neural Networks

Neural networks are brain-inspired systems deigned to mimic how humans learn. They learn and increase their accuracy over time by using training data. Although neural networks are mostly suitable for large and highly complex machine learning tasks, it yielded positive outcomes in this model.

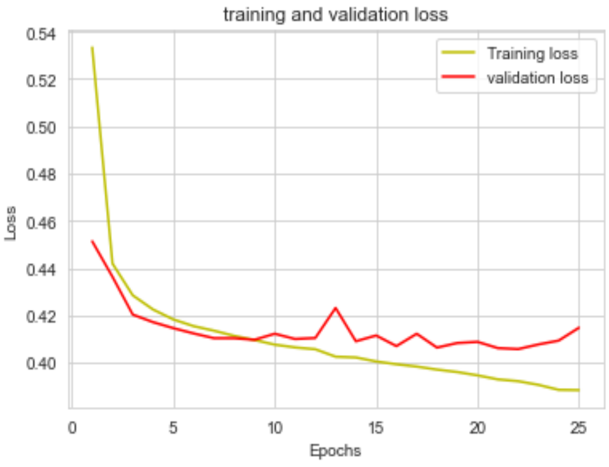


Figure 5. Graph of Validation and Training Loss.

VII. Conclusion

This study evaluates the performance of six machine learning algorithms in predicting customer churn in a telecom company. The data used in this study is open source and complies with relevant data protection policies. The modelling was done without bias with all professional and ethical issues considered. Although all the models performed better than the baseline model, Random Forest Classifier performed best with an accuracy of 84% and AUROC of 0.920. This study can further be carried out with a more robust dataset with the aim of creating an application to predict customer churn in real time.

VIII. References

[1] Customer Acquisition and Retention. Available at: https://www.invespcro.com/blog/customer-acquisition-retention/ .

[2] Friedman, J.H. (2001) 'Greedy function approximation: A gradient boosting machine', The Annals of Statistics, 29(5), pp. 1189-1232. doi: 10.1214/aos/1013203451.

[3] Geron, A. (2019) Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Sebastopol: O'Reilly Media, Incorporated.

[4] Mahajan, V., Misra, R. and Mahajan, R. (2017) 'Review on factors affecting customer churn in telecom sector', International Journal of Data Analysis Techniques and Strategies, 9(2), pp. 122-144. doi: 10.1504/IJDATS.2017.085898.

[5] Reichheld, F.F. and Sasser, J., W E (1990) 'Zero defections: quality comes to services', Harvard Business Review, 68 (5), pp. 105.

[6] (2021) 'Telecom Global Market Report 2021: COVID 19 Impact and Recovery to 2030', NASDAQ OMX's News Release Distribution Channel, Feb 2,.

[7] Tianyuan Zhang, Sérgio Moro and Ricardo F Ramos (2022) 'A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation', Future Internet, 14(3), pp. 94. doi: 10.3390/fi14030094.

[8] Tin Kam Ho (1995) 'Random decision forests', - Proceedings of 3rd International Conference on Document Analysis and Recognition. doi: 10.1109/ICDAR.1995.598994.

[9] X. Zhang, G. Feng and H. Hui (2009) 'Customer-Churn Research Based on Customer Segmentation', - 2009 International Conference on Electronic Commerce and Business Intelligence. doi: 10.1109/ECBI.2009.86.

[10] Ke, G. et al. (2017) 'LightGBM: A Highly Efficient Gradient Boosting Decision Tree', Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA, Curran Associates Inc.

[11] Understanding the ROC curve and AUC. Available at: https://towardsdatascience.com/understanding-the-roc-curve-and-auc-dd4f9a192ecb .

IX. Appendix

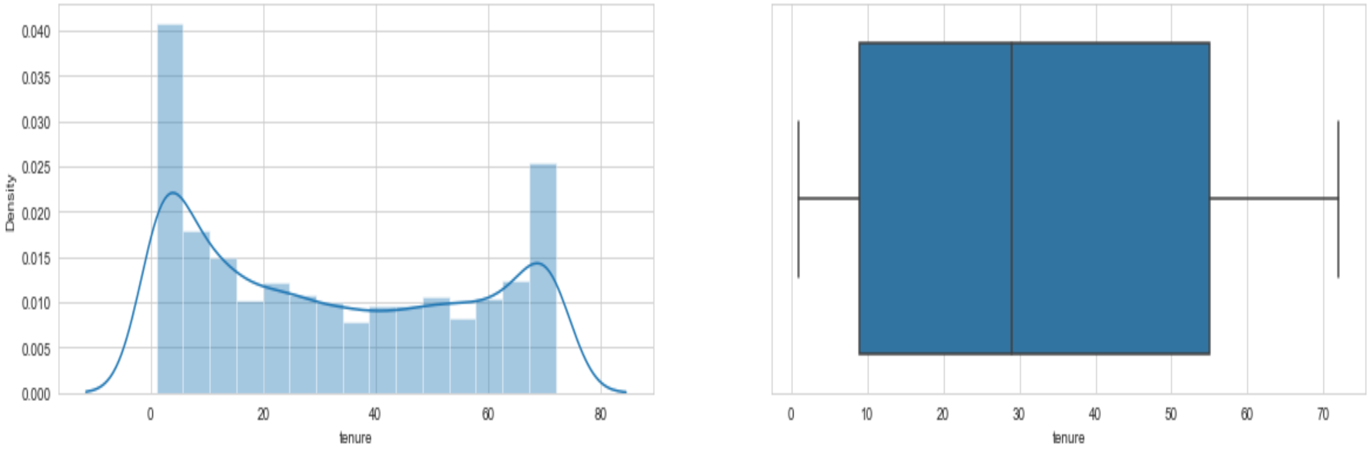


Figure 6. distribution of Tenure Column

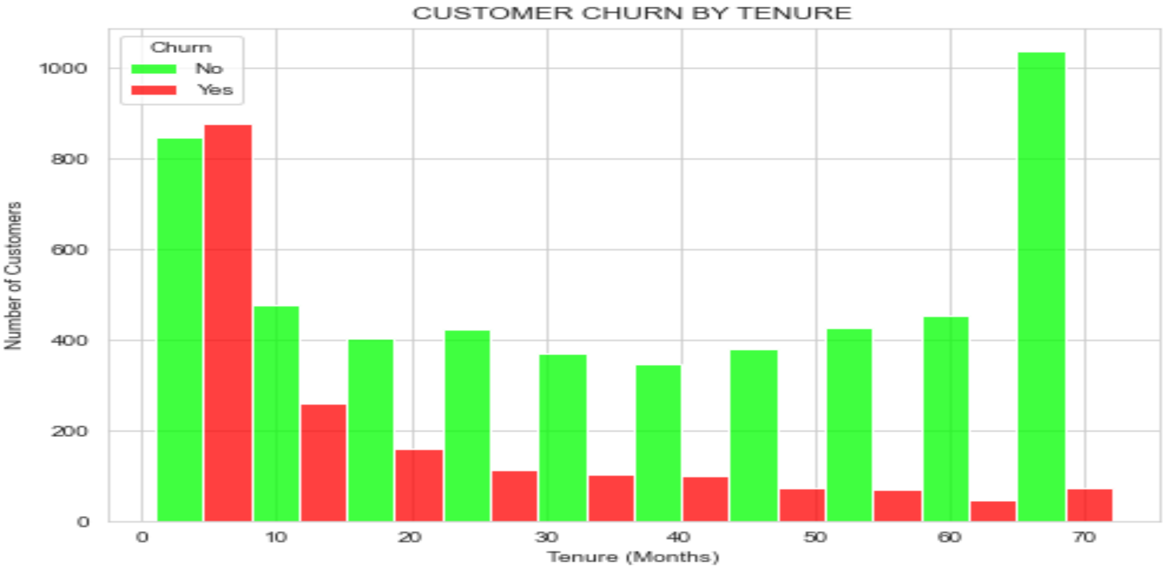


Figure 7

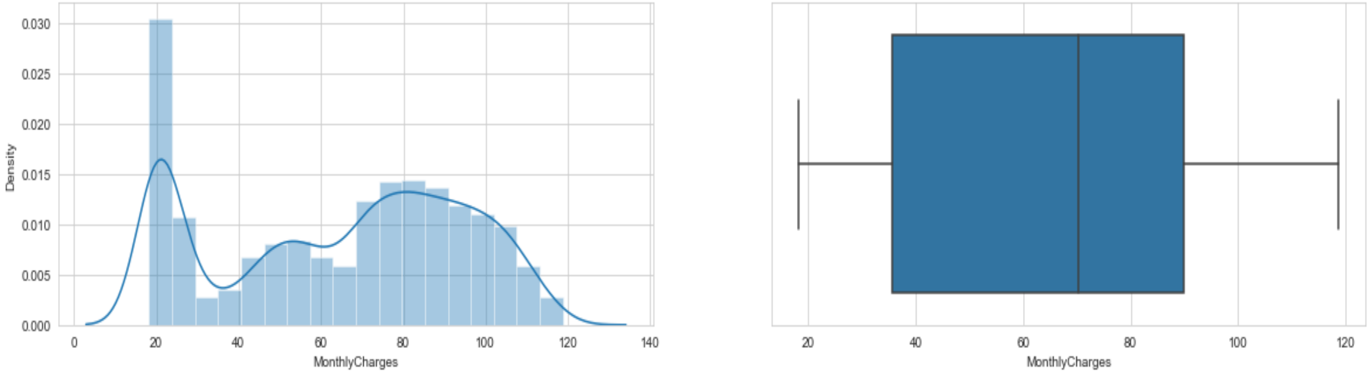


Figure 8. distribution of monthly charges column

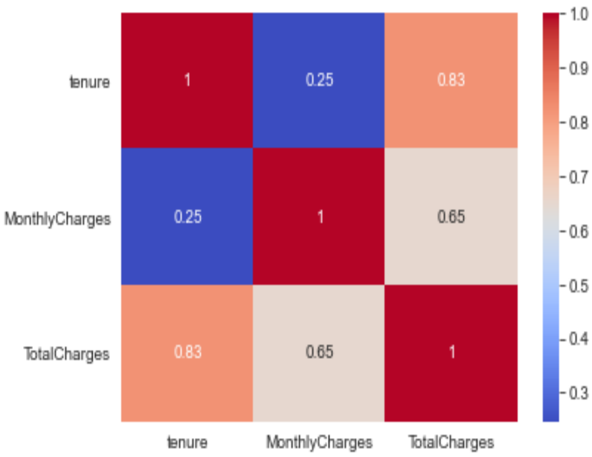


Figure 9. check for collinearity

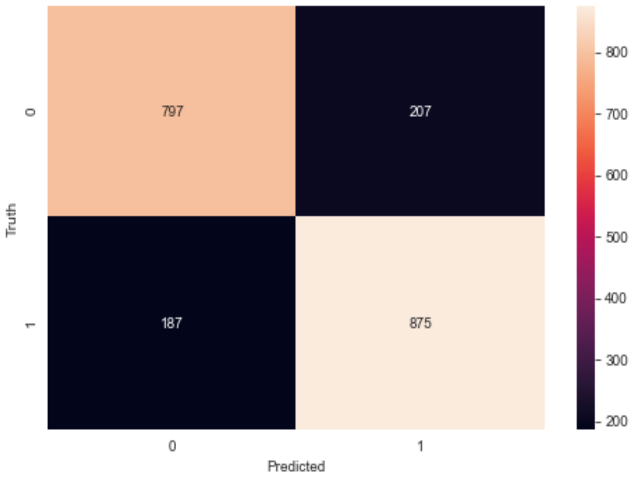


Figure 10 Logistic Regression Confusion Matrix

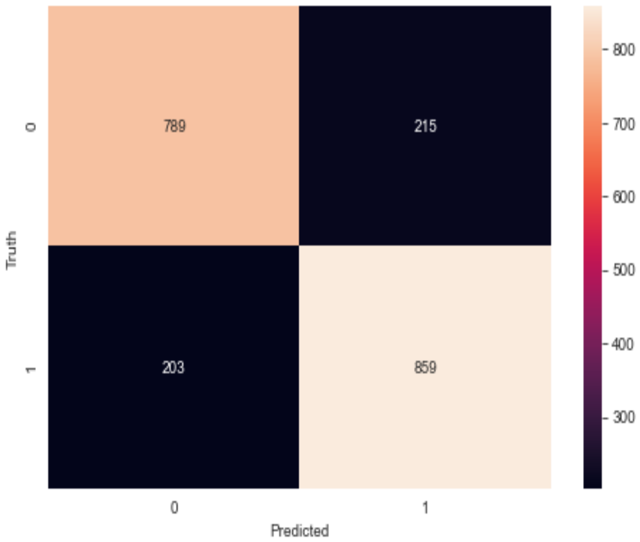


Figure 11 Decision Tree Confusion Matrix

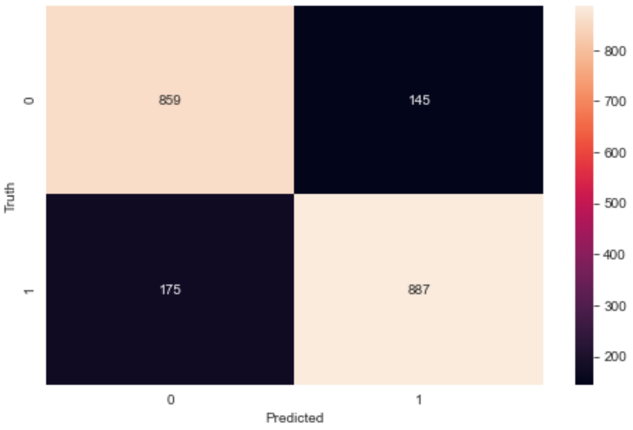


Figure 12. Random Forest confusion matrix

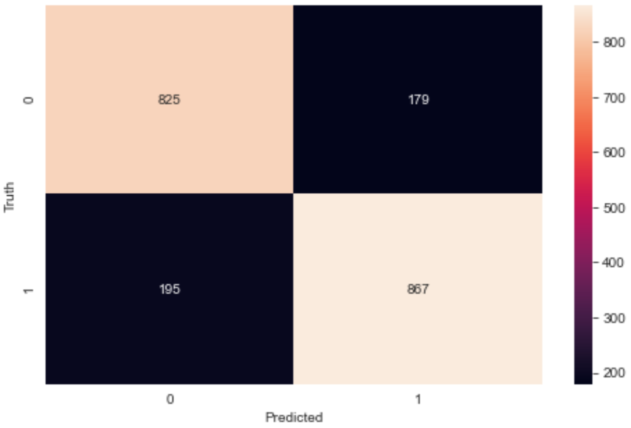


Figure 13. SVC Confusion Matrix

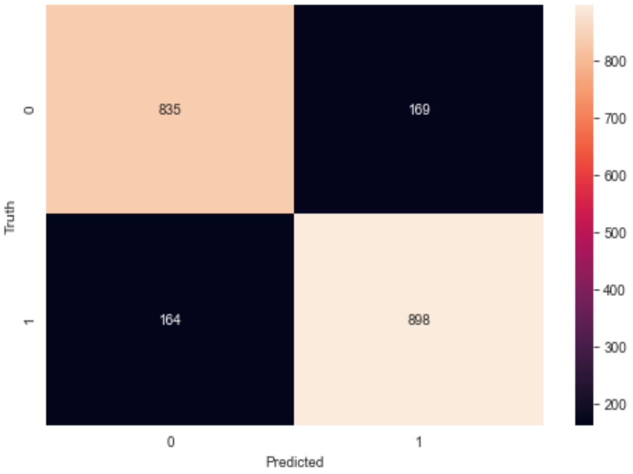


Figure 14. LightGBM Confusion matrix

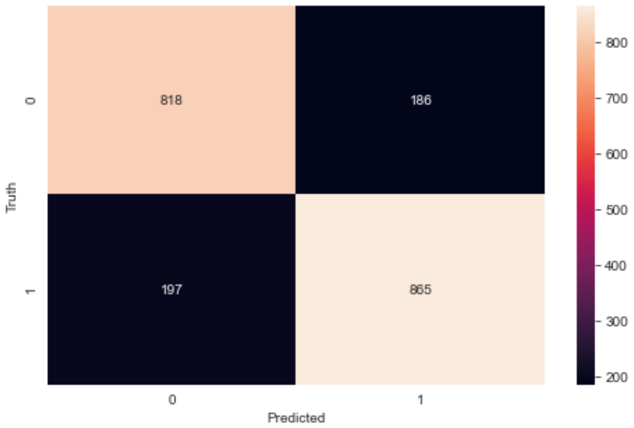


Figure 15. ANN confusion matrix

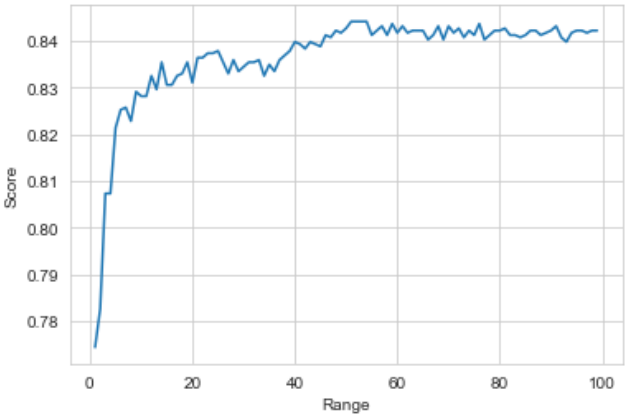


Figure 16 Plot of best estimator for Random Forest