Predicting Customer Churn in a Telco Company using Demographic and Service Data

Justine George(JXG210092), Harshavardhini Sridhar(HXS220004)

Abstract

This project aims to predict customer churn in a telecommunications company using a dataset containing demographic and service data. The dataset, sourced from Kaggle, consists of 7043 customers and 21 features. Our approach involved performing exploratory data analysis, addressing class imbalance through synthetic minority oversampling (SMOTE) and edited nearest neighbor (ENN) techniques, and evaluating various machine learning algorithms such as XGBoost, KNN, SVM, Random Forest, and more. The models were assessed using cross-validation techniques and various performance metrics, including precision, recall, F1 score, and area under the ROC curve. Our findings demonstrated that certain models performed better than others in predicting churn, providing valuable insights for the telecommunication industry to implement targeted customer retention strategies.

Introduction

Customer churn, or the loss of customers. Customer churn can lead to considerable revenue losses and reduced market share. Retaining existing customers is often more cost-effective than acquiring new ones, which is why accurately predicting and addressing customer churn is crucial for companies to maintain profitability and customer satisfaction.

Dataset Description

The dataset used for this project is obtained from Kaggle, originally sourced from IBM Sample Data Sets. The dataset contains information on 7043 customers, with each customer represented by 21 columns (features).

Exploratory Data Analysis

For visualization purposes we have divided the features into three categories (plots are available in notebook) :

- Group 1 : Customer Information :('gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure')
- Group 2 : Services signed up for: ('PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies')
- Group 3 : Payment ('Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges')

Each feature is analysed with Churn variable to understand their relationship, example:

- Senior citizens tend to have slightly higher monthly charges.
- Customers who have a partner are more likely to have dependents.
- Customers who have been with the company for a longer time are more likely to have a partner or dependents.
- Customers who have phone service tend to have higher monthly charges.

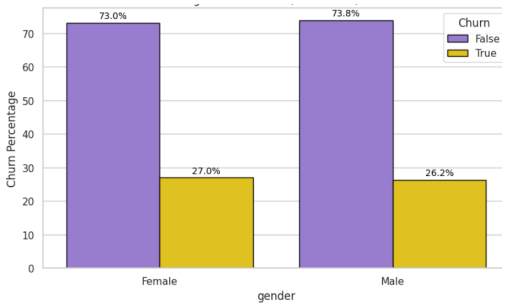


Figure 1: Gender vs Churn

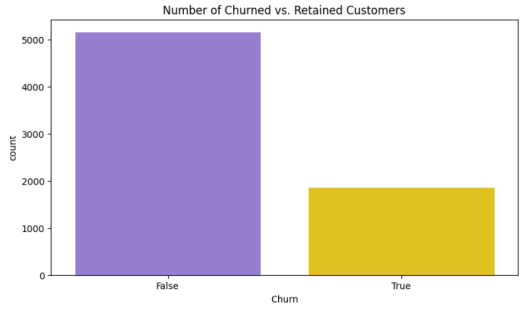


Figure 2: Number of Churned vs Retained Customers

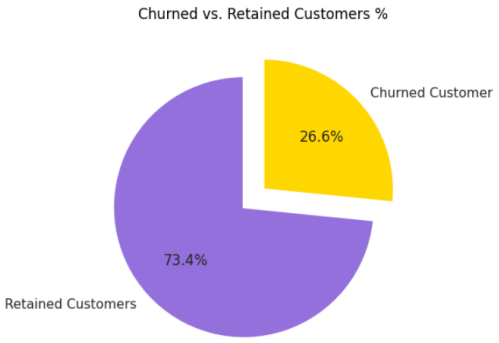


Figure 3: Churned vs Retained Customers

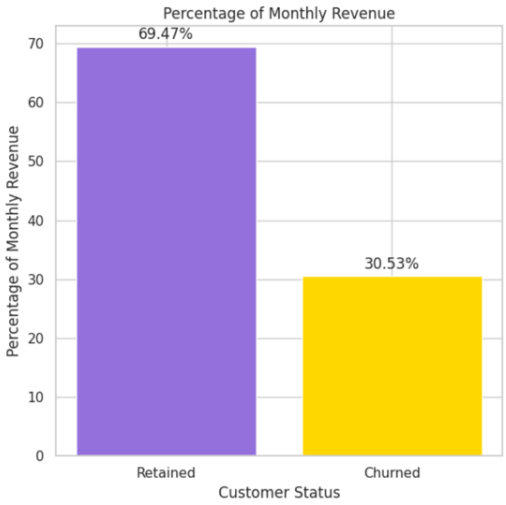


Figure 4: Total monthly Revenue loss-30 percent revenue loss due to customer churn

- Customers who use paperless billing tend to have higher monthly and total charges.
- Customers who have higher monthly charges tend to accumulate higher total charges over time.
- Customers who have been with the company for a longer time are less likely to churn.
- Customers who have higher total charges are slightly less likely to churn.

Feature Selection

On performing Chi Square test, we find that Gender and PhoneService do not significantly affect Churn, as indicated by their P-values (¿ 0.05). Therefore, these features can be dropped from the analysis.

Data Preprocessing

To tackle class imbalance, we take the following steps: One-hot encode categorical features, dropping the first category of each feature to avoid multicollinearity. Split the dataset into training and testing sets (80:20 ratio) for evaluating model performance on unseen data. Apply Synthetic Minority Over-sampling Technique (SMOTE) to oversample the minority class (churning customers) and balance the target variable ('Churn') in the training set. Apply Edited Nearest Neighbor (ENN) technique to remove noisy samples from the majority class (non-churning customers).

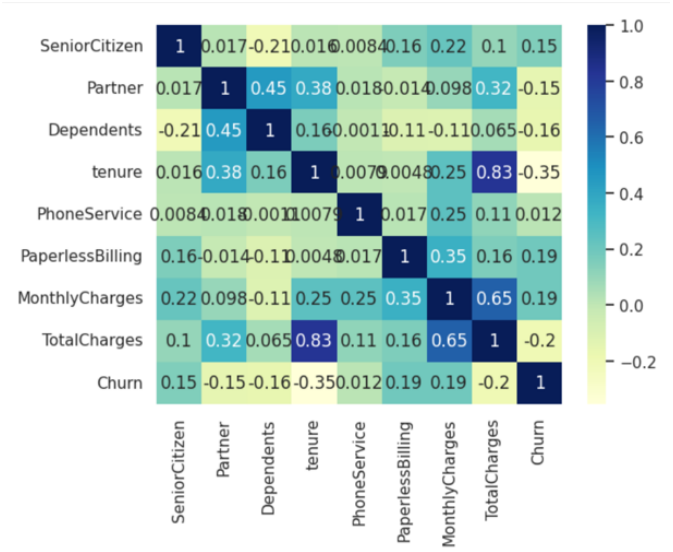


Figure 5: Correlation Matrix Heatmap

Feature Scaling

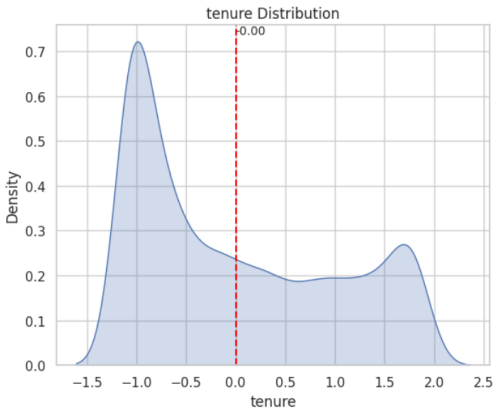


Figure 6: Tenure distribution

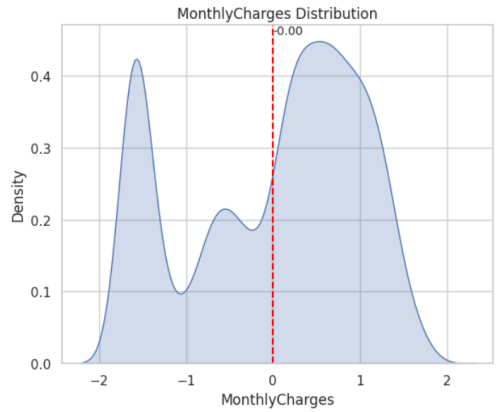


Figure 7: Monthly Charges distribution

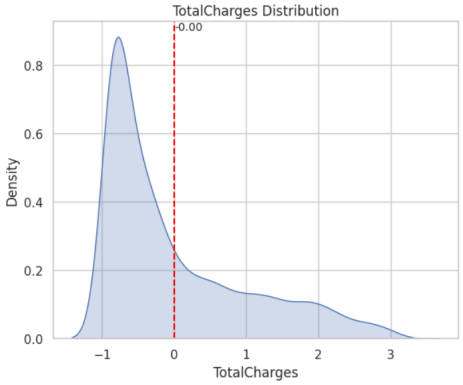


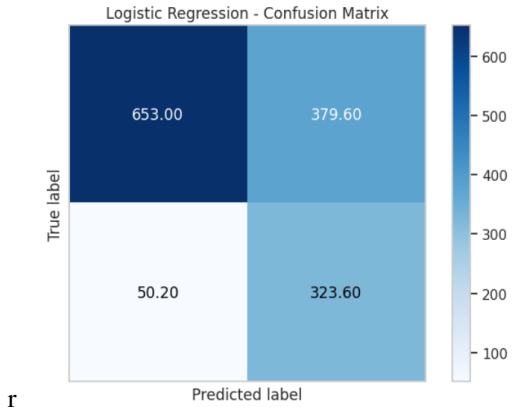
Figure 8: Total Charges Distribution

Continuous features such as tenure, MonthlyCharges, and TotalCharges are scaled using standard scaling to ensure that they contribute equally to the model. This helps the model converge faster and avoids any one feature from dominating the other features. We apply standard scaling to these features, transforming them to have a mean of 0 and a standard deviation of 1.

Machine Learning models using skLearn

Logistic regression

- Why was this algorithm chosen ? Logistic Regression is a common and popular algorithm for binary classification problems like churn prediction. It's useful when the relationship between the input features and the output variable is linear and in our case, many features have a linear relationship with Churn.
- How does this algorithm perform for our dataset? It performs reasonably well, the true positive rate (sensitivity) of 0.865 indicates the model is able to correctly identify a large proportion of customers who are likely to churn, and the false positive rate of 0.367 indicates that the model has a relatively high rate of false positives.
- Why was the performance moderate? Logistic Regression does not perform well in datasets (eg. telco customer churn) with non linear data. It also assumes that features



r

Figure 9: Confusion Matrix of Logistic Regression

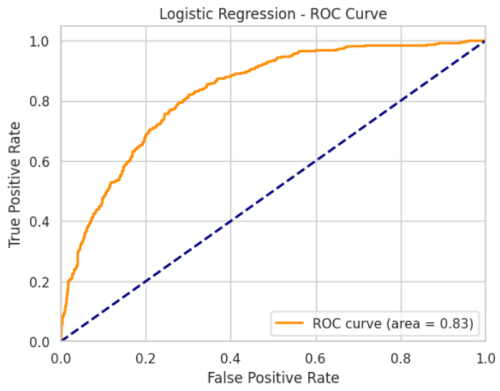


Figure 10: ROC Curve of logistic regression

are independent of each other but in reality, features are correlated.

Stochastic Gradient Descent Classifier

- Why was this algorithm chosen ? SGD classifier is widely used for classification problems. Our dataset contains many records and columns so we need a fast and computationally efficient algorithm like Stochastic Gradient Descent algorithm which can handle large datasets.
- How does this algorithm perform for our dataset? The evaluation metrics value suggest that the SGD classifier has moderate performance in predicting churn. The TPR 0.75 is relatively high, indicating that the classifier is good at identifying customers who are likely to churn, and the FPR 0.37 is relatively high. The F1 score indicates that the classifier's precision and recall are not well-balanced. The area of 0.78 suggests that the classifier is able to correctly classify around 0.78 of the samples as either positive or negative.
- Why was the performance moderate? SGD is a linear model and may not be able to capture the non-linear relationships and the feature interactions effectively, leading to poor performance.

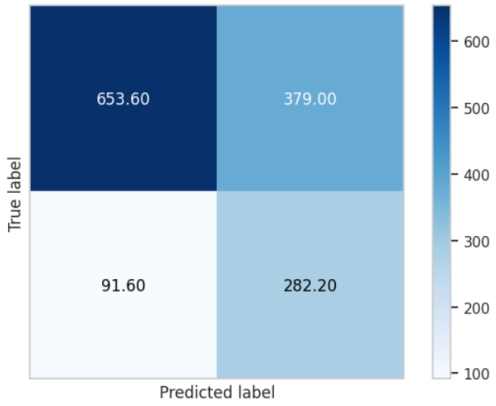


Figure 11: Confusion Matrix of SGD Classifier

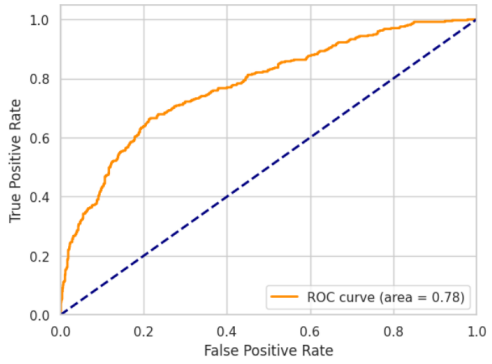


Figure 12: ROC Curve of SGD Classifier

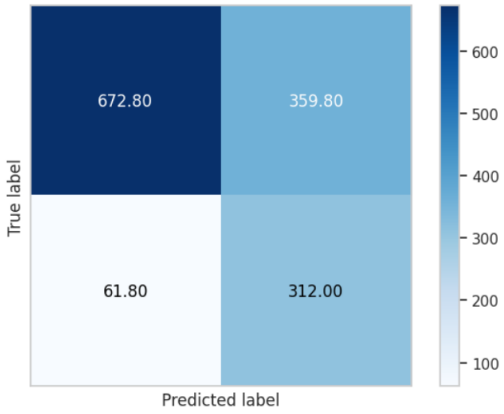


Figure 13: Confusion Matrix of XGBoost

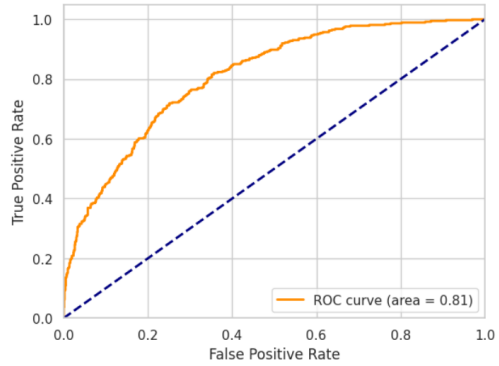


Figure 14: ROC Curve of XGBoost

XGBoost

- Why was this algorithm chosen ? XGBoost has built-in mechanisms to handle imbalanced data, such as weighted and subsample options, which can help to improve the model's performance. XGBoost is good at feature selection which is helpful for datasets with a large number of variables.
- How does this algorithm perform for our dataset? An accuracy of 0.7 means that the model is correct 70 percent of the time. The high TPR of 0.83 indicates that the model is good at correctly identifying customer churns which is an important component in our project. FPR of 0.35 indicates that it can be reduced so that the company does not unnecessarily offer incentives to customers who are unlikely to churn.
- Why was the performance moderate? The data available is insufficient for the model to accurately learn and make accurate predictions since there are some complex relationships in telco customer churn dataset that the XGBoost classifier is unable to learn.

KNN

- Why was this algorithm chosen ? Knn algorithm is simple and fast and our dataset contains many records and features to be considered. It can also capture complex and nonlinear relationships between the features and the target variable.
- How does this algorithm perform for our dataset? Although the TPR is quite high, the accuracy is low because of high FPR. A high FPR can be costly for a telco company. 0.537 F1 score indicates that the model has moderate performance in terms of precision and recall.
- Why was the performance moderate? Knn does not perform well for our dataset due to curse of dimensionality.

SVM

- Why was this algorithm chosen ? SVM is a commonly used ML algorithm for classification tasks.It models non linear decision boundaries and can handle high dimensional data without overfitting.
- How does this algorithm perform for our dataset? The performance of SVM is not good. The accuracy is too low. Although the TPR is high which can be a good thing, the FPR is also high which can cause unnecessary costs to the company.
- Why was the performance moderate? We did not choose appropriate model parameters such as the kernel function, regularization parameter, and gamma value that is specific

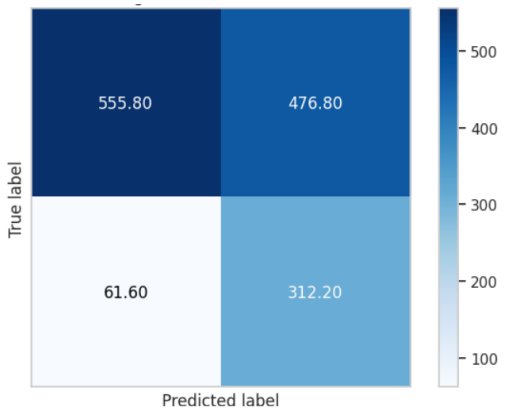


Figure 15: Confusion Matrix of KNN

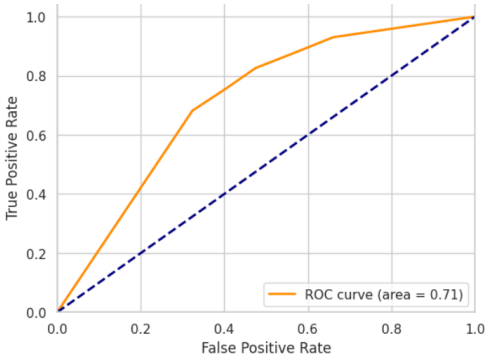


Figure 16: ROC Curve of KNN

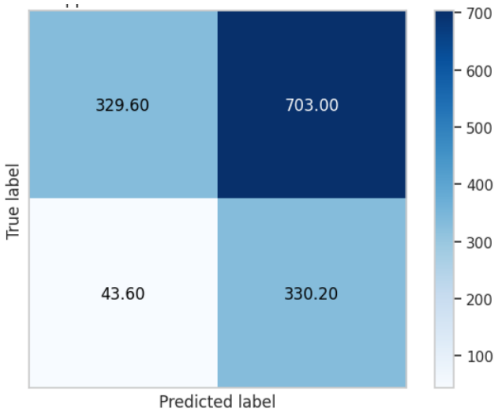


Figure 17: Confusion Matrix of SVM

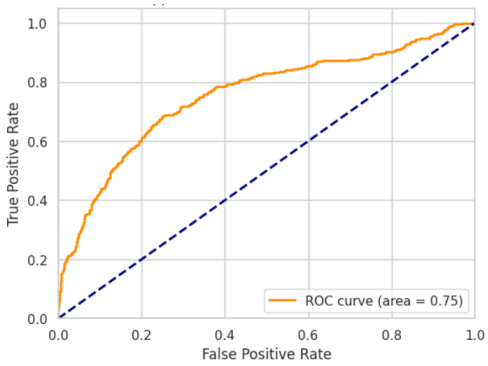


Figure 18: ROC Curve of SVM

to our dataset, this is the reason why we did not get accurate results. Addressing this issue can improve the model performance.

Random forest

- Why was this algorithm chosen ? Random forests can capture non linear relationships that are present in the customer churn dataset and it also provides feature importance rankings.
- How does this algorithm perform for our dataset? The model exhibits a not so bad accuracy with a high TPR of 0.84 and this indicates that the model effectively predicts the customers who will churn. The FPR is comparable to other models we have used. F1 score is relatively better than the above models.
- Why was the performance moderate? The hyperparameters like number of trees, maximum depth of each tree, number of fetures at each split has not been provides that is specific to our dataset and these hyperparamters play a big role in the performance of Random forests.

Decision tree

- Why was this algorithm chosen ? Decision trees can handle both categorical and numerical data and it can capture non linear data in the telco customer churn dataset.
- How does this algorithm perform for our dataset? An accuracy of 0.69 indicates that the decision tree algorithm correctly predicts 69percent of the time, if the customer will churn or not. The TPR is high and the FPR value is similar to other model's outputs.
- Why was the performance moderate? Since we have many features in the dataset and decision trees suffer from curse of dimensionality, the tree struggles to find meaningful splits.

Naive Bayes

- Why was this algorithm chosen ? Naive Bayes algorithm is a simple and efficient classification algorithm since it requires minimal computation. The algorithm depends on very less hyperparamters.
- How does this algorithm perform for our dataset? An accuracy of 0.68 indicates that the decision tree algorithm correctly predicts if the customer will churn or not by 68 percent. The TPR is high and the FPR value is similar to other model's outputs.
- Why was the performance moderate? Naive Bayes assumes that the features are conditionally independent of each other given the class label. But in telco customer churn dataset, the features are highly correlated.

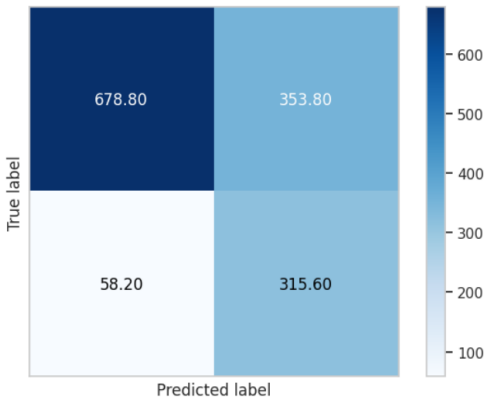


Figure 19: Confusion Matrix of Random forest

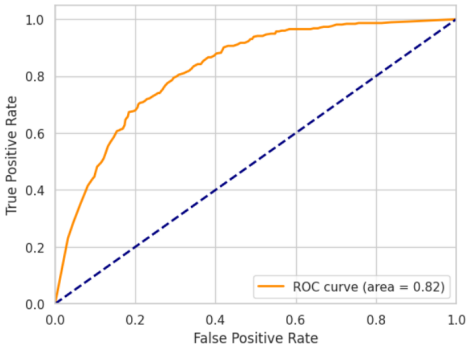


Figure 20: ROC Curve of Random forest

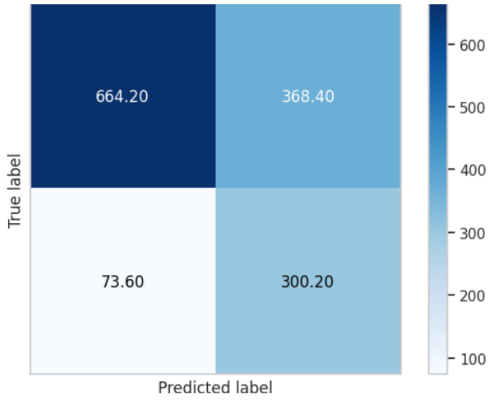


Figure 21: Confusion Matrix of Decision Tree

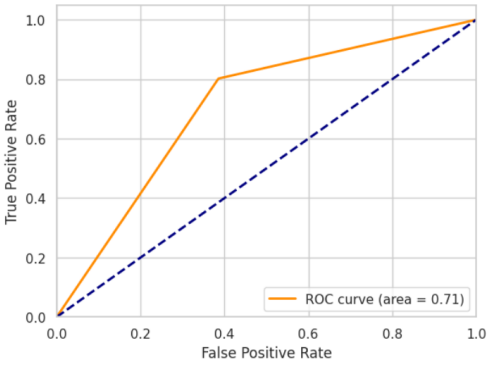


Figure 22: ROC Curve of Decision tree

Kernel SVM

Confusion Matrix and ROC Curve of Kernel SVM can be found in the notebook

- Why was this algorithm chosen ? Kernel SVM can model non linear decision boundaries and this is useful as there is non linear relationship between customer related features and churn.
- How does this algorithm perform for our dataset? Although there is a high TPR of 0.88, the accuracy is too low i.e 0.46 because of high FPR of 0.68.
- Why was the performance moderate? Even after doing SMOTE and ENN, the class imbalance problem is not fully resolved.

Artificial Neural Network (ANN)

Confusion Matrix and ROC Curve of ANN can be found in the notebook

- Why was this algorithm chosen ? ANN can capture the non linear relationships in the telco customer churn dataset and can recognise complex patterns.
- How does this algorithm perform for our dataset? The TPR is high with 0.82 and the FPR is also high (0.32) which means the model has large number of false alarms.
- Why was the performance moderate? We have not customized the ANN with hyperparameters like activation function, regularization method that can optimise model performance.

Model Comparison and Conclusion

- ANN gives the best accuracy.
- KNN and Kervel SVM have a high TPR
- ANN has the lowest FPR
- ANN has the highest F1 score

Overall performance of ANN is better than all other algorithms. This is because, ANN captures the complex non-linear relationships between features and Churn.

References

[1] Telco Customer Churn Dataset - Kaggle https://www.kaggle.com/datasets/blastchar/telco-customer-churn
[2] IBM Business Analytics - Customer Churn Dataset
[3] Telco Customer Churn Prediction - Kaggle https://www.kaggle.com/code/gaganmaahi224/telco-customer-churn-prediction-with-11-ml-algos

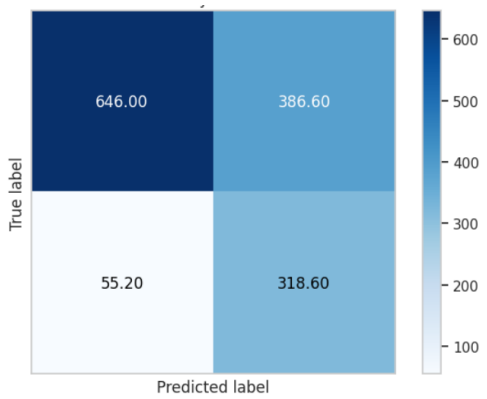


Figure 23: Confusion Matrix of Naive Bayes

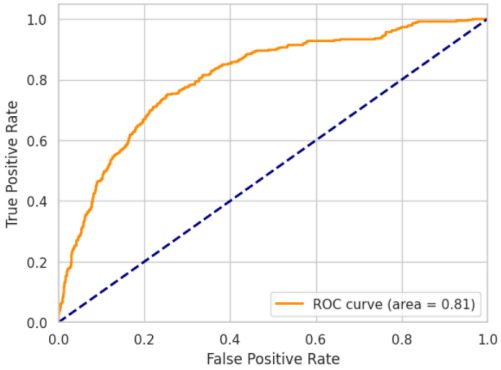


Figure 24: ROC Curve of Naive Bayes

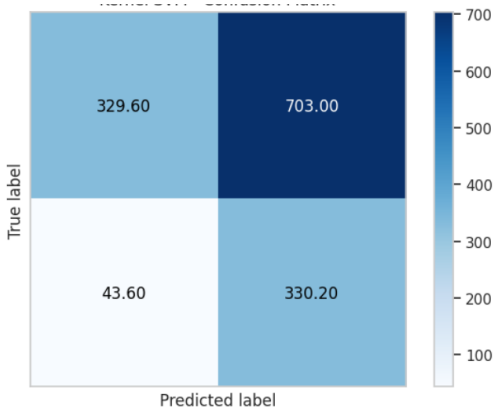


Figure 25: Confusion Matrix of Kernel SVM

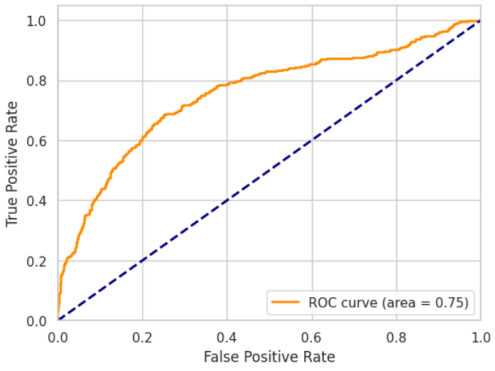


Figure 26: ROC Curve of Kernel SVM

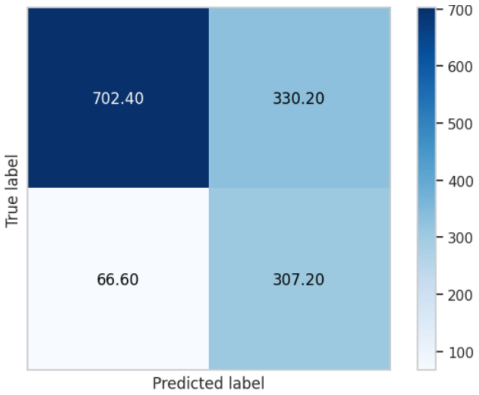


Figure 27: Confusion Matrix of ANN

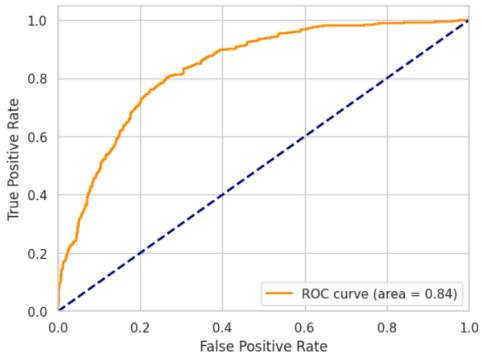


Figure 28: ROC Curve of ANN

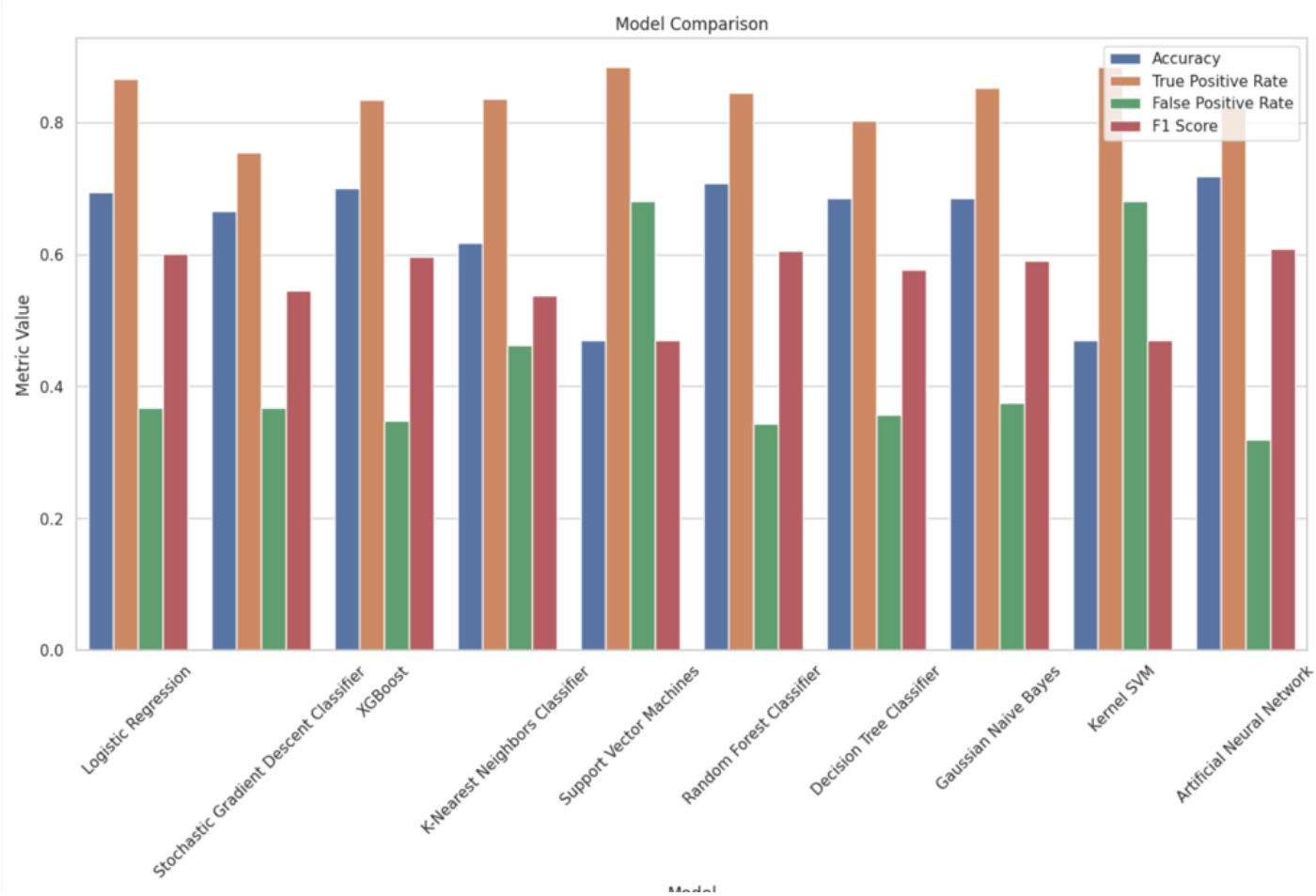


Figure 29: Comparison of all models