CS 513: KNOWLEDGE DISCOVERY AND DATA MINING

Customer Churn Analysis

Team 2

Pranit Dutta
20010681

Sushil Rajeeva Bhanary
20015528

Sabitha Rachel Nazareth
20012150

Nihal Sanjay Palled
20011136

Problem Statement

- Background: Customer churn is a significant concern for telecommunications companies, as retaining existing customers is often more cost-effective than acquiring new ones.
- The objective of this project is to predict customer churn in a telecommunications company using a dataset containing customer information and historical data. By developing a predictive model, we aim to identify the most significant factors contributing to customer churn and classify customers as high or low risk of churning. This will enable the company to take proactive measures to retain customers and improve customer satisfaction.

Dataset Explanation

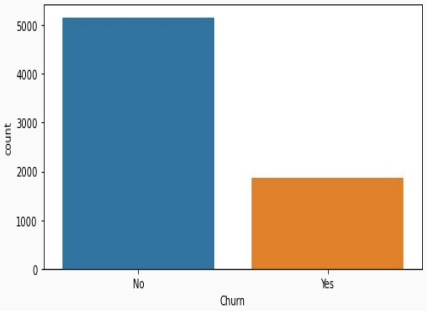
Field	Description
customerID	A unique identifier for each customer
gender	Customer's gender (Male or Female)
SeniorCitizen	Whether the customer is a senior citizen (1) or not (0)
Partner	Whether the customer has a partner (Yes or No)
Dependents	Whether the customer has dependents (Yes or No)
tenure	The number of months the customer has stayed with the company
PhoneService	Whether the customer has phone service (Yes or No)
MultipleLines	Whether the customer has multiple lines (Yes, No, or No phone service)
InternetService	Customer’s internet service provider (DSL, Fiber optic, or No)
OnlineSecurity	Whether the customer has online security (Yes, No, or No internet service)

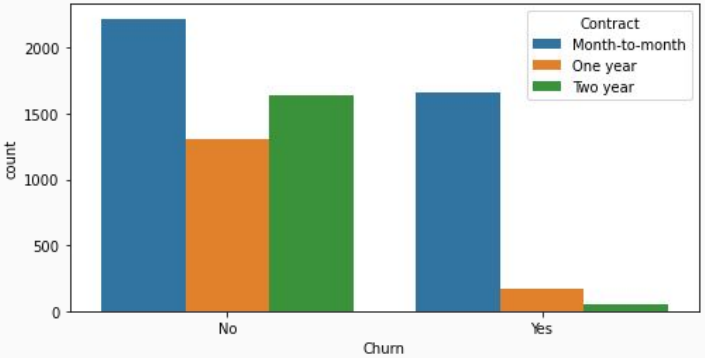
Dataset Explanation

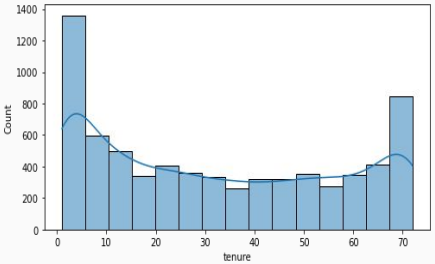
Field	Description
OnlineBackup	Whether the customer has online backup (Yes, No, or No internet service)
DeviceProtection	Whether the customer has device protection (Yes, No, or No internet service)
TechSupport	Whether the customer has tech support (Yes, No, or No internet service)
StreamingTV	Whether the customer has streaming TV (Yes, No, or No internet service)
StreamingMovies	Whether the customer has streaming movies (Yes, No, or No internet service)
Contract	The contract term of the customer (Month-to-month, One year, or Two year)
PaperlessBilling	Whether the customer has paperless billing (Yes or No)
PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), or Credit card (automatic))
MonthlyCharges	The amount charged to the customer monthly
TotalCharges	The total amount charged to the customer
Churn	Whether the customer churned (Yes or No) - This is the target variable

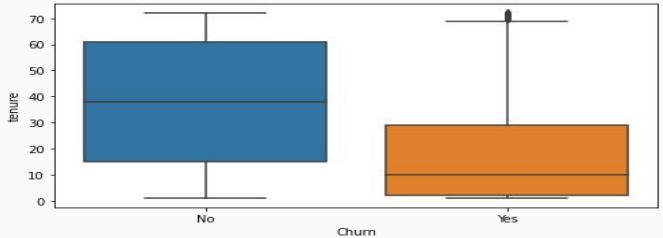
Data Processing

- The data was preprocess by converting the non numerical object to integer objects by using label encoder function.
- During the EDA Phase of the project we observed that there is an unbalance in the data set. The ratio of the churn to no churn was in the ratio of 3:1 . In order to handle the unbalance we used sampling technique called SMOTE-ENN.
- Then feature scaling called Min-Max was applied

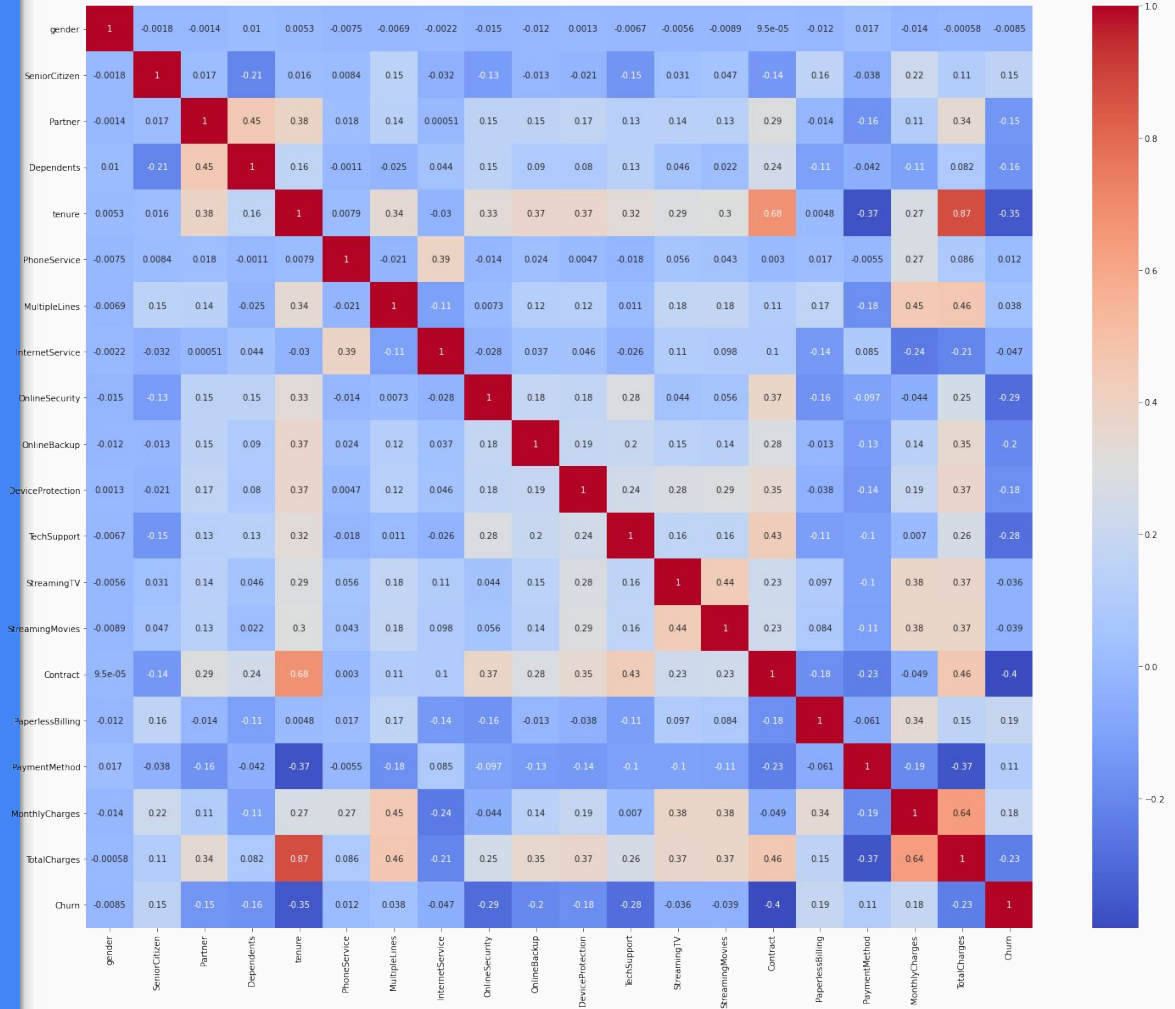








Correlation Matrix



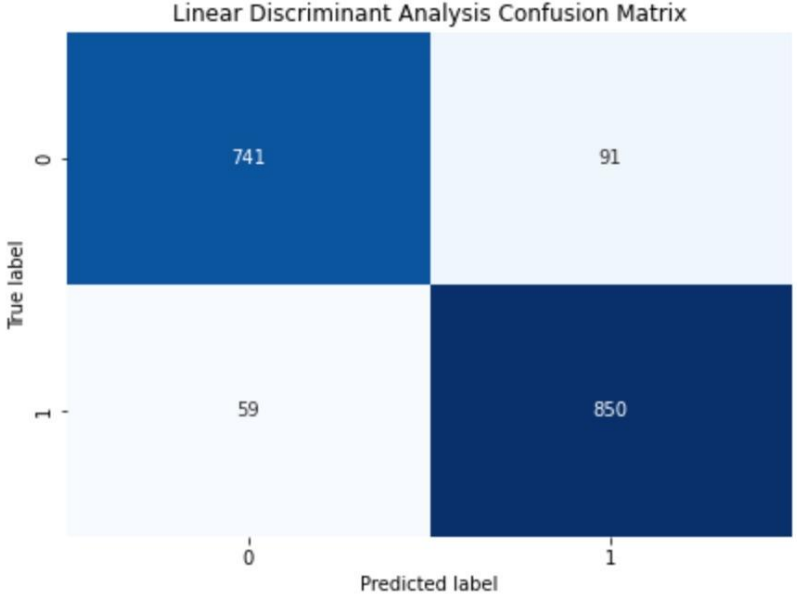
MODELS

- LDA
- XGboost
- SVM
- ADA BOOST
- Neural Network
- Neural Network with grid search
- Label prop
- Naive bayes
- KNN
- Random Forest
- Decision Tree
- Logistic Regression
- Bagging Classifier

Metrics

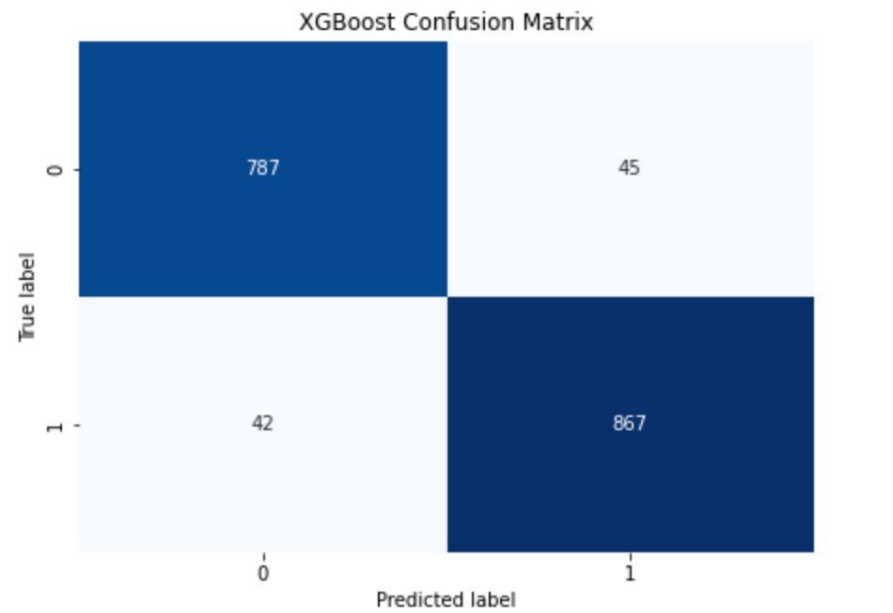
- Accuracy
- Balanced Accuracy
- ROC AUC score
- F1 score
- Precision
- Recall
- Cohen’s Kappa

Linear Discriminant Analysis



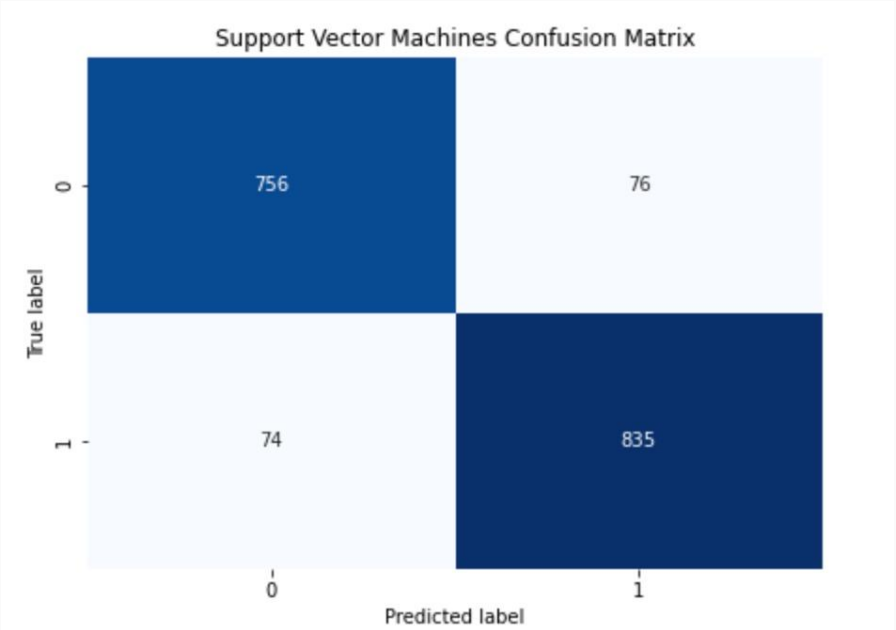

```
Accuracy:  0.9138426191843768
Precision:  0.9032943676939427
Recall:  0.935093509350935
F1:  0.918918918918919
Balanced Accuracy:  0.9128592546754675
ROC AUC score:  0.9128592546754675
Cohens Kappa:  0.8270658070394779
```

XGBoost



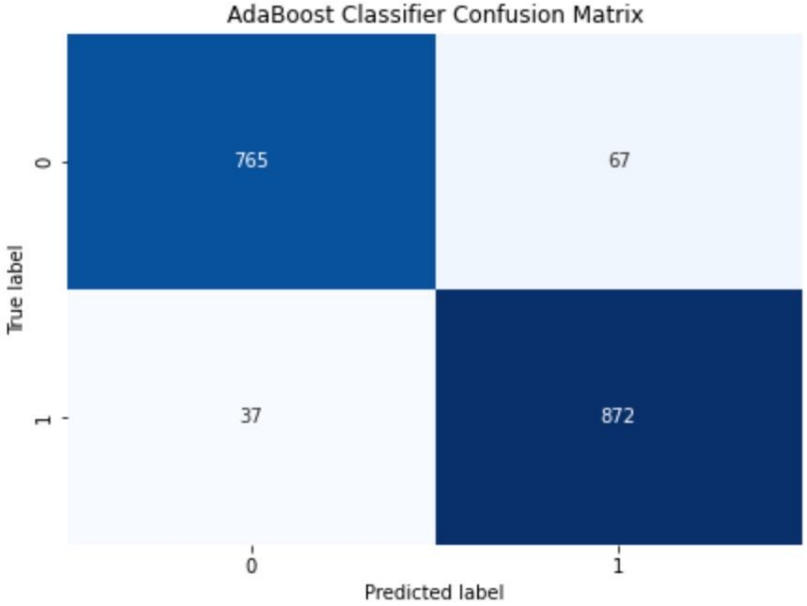

```
Accuracy:  0.9500287191269385
Precision:  0.9506578947368421
Recall:  0.9537953795379538
F1:  0.9522240527182867
Balanced Accuracy:  0.9498544205382077
ROC AUC score:  0.9498544205382078
Cohens Kappa:  0.8998462652370987
```

Support Vector Machine



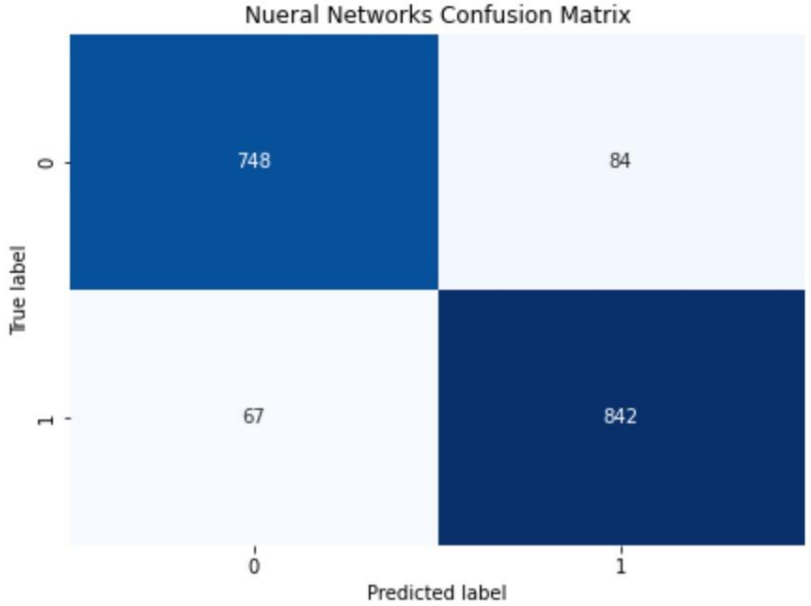

```
Accuracy:  0.9138426191843768
Precision:  0.9165751920965971
Recall:  0.9185918591859186
F1:  0.9175824175824175
Balanced Accuracy:  0.9136228526698824
ROC AUC score:  0.9136228526698823
Cohens Kappa:  0.8273299383373159
```

ADA Boost




```
Accuracy:  0.9402642159678346
Precision:  0.9286474973375932
Recall:  0.9592959295929593
F1:  0.9437229437229436
Balanced Accuracy:  0.9393835417195566
ROC AUC score:  0.9393835417195565
Cohens Kappa:  0.8801111857116561
```

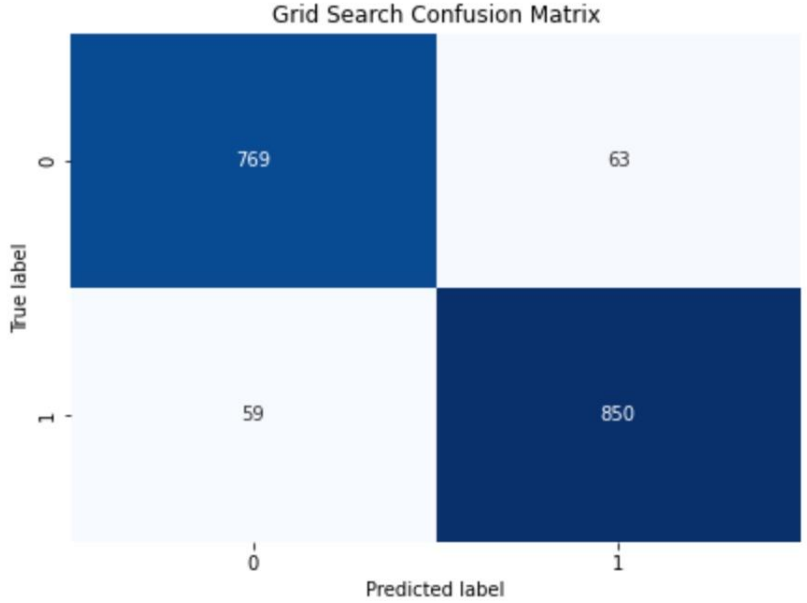
Neural Network




```
Accuracy:  0.913268236645606
Precision:  0.9092872570194385
Recall:  0.9262926292629263
F1:  0.9177111716621253
Balanced Accuracy:  0.912665545400694
ROC AUC score:  0.9126655454006939
Cohens Kappa:  0.8260459601116149
```

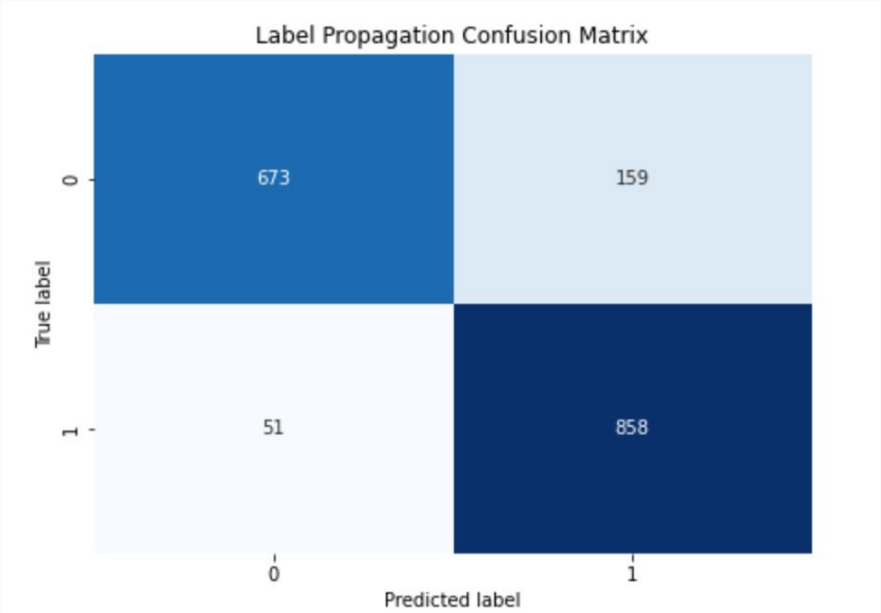
Neural Network With Grid Search

```
{'activation': 'relu', 'solver': 'lbfgs'}
```



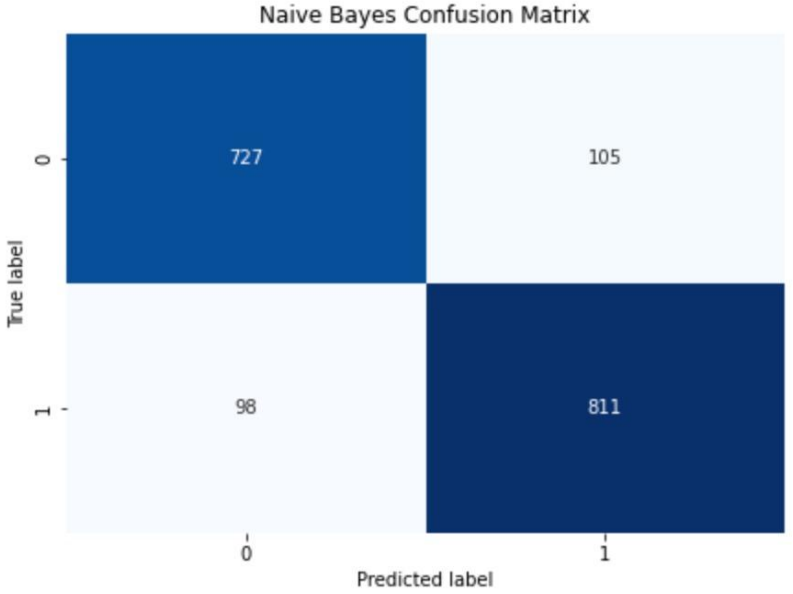

```
Accuracy:  0.9299253302699598
Precision:  0.9309967141292442
Recall:  0.935093509350935
F1:  0.9330406147091108
Balanced Accuracy:  0.9296861777523906
ROC AUC score:  0.9296861777523905
Cohens Kappa:  0.8595473818132765
```

Label Propagation



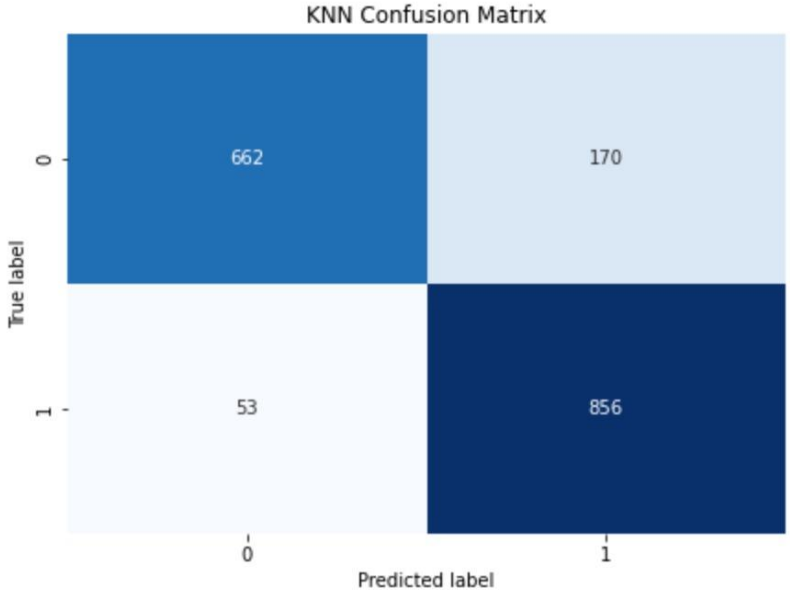

```
Accuracy:  0.8793796668581275
Precision:  0.8436578171091446
Recall:  0.9438943894389439
F1:  0.8909657320872274
Balanced Accuracy:  0.8763943101040873
ROC AUC score:  0.8763943101040875
Cohens Kappa:  0.7569502612580272
```

Naive Bayes



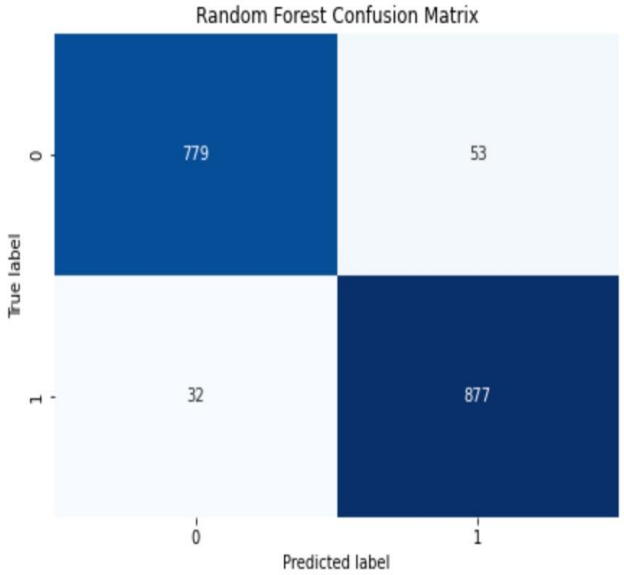

```
Accuracy:  0.8834003446295232
Precision:  0.8853711790393013
Recall:  0.8921892189218922
F1:  0.8887671232876712
Balanced Accuracy:  0.8829936479224845
ROC AUC score:  0.8829936479224846
Cohens Kappa:  0.7662603494491207
```

K-Nearest Neighbours



Accuracy: 0.8719126938541069
Precision: 0.834307992202729
Recall: 0.9416941694169417
F1: 0.8847545219638244
Balanced Accuracy: 0.8686836231700092
ROC AUC score: 0.8686836231700092
Cohens Kappa: 0.741785367728874

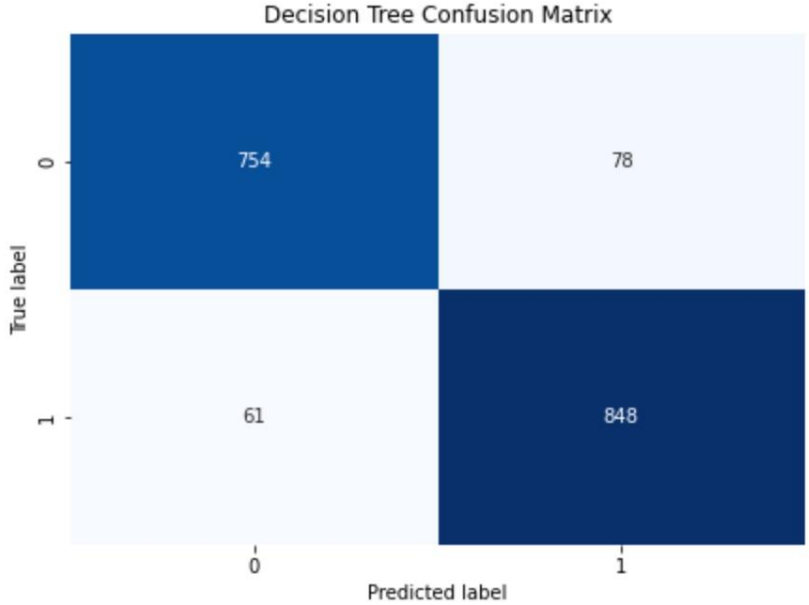
Random Forest



Accuracy: 0.9511774842044802
Precision: 0.943010752688172
Recall: 0.9647964796479648
F1: 0.9537792278412179
Balanced Accuracy: 0.9505472782855209
ROC AUC score: 0.9505472782855207
Cohens Kappa: 0.9020588910751384

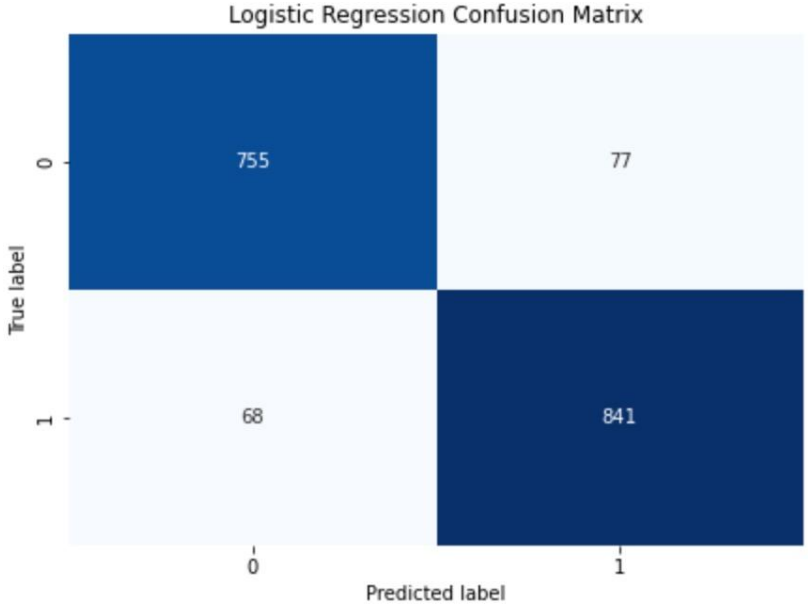
	feature	importance
0	gender	0.009921
1	SeniorCitizen	0.005474
2	Partner	0.028714
3	Dependents	0.022435
4	tenure	0.167998
5	PhoneService	0.003993
6	MultipleLines	0.009282
7	InternetService	0.054473
8	OnlineSecurity	0.117993
9	OnlineBackup	0.039395
10	DeviceProtection	0.018390
11	TechSupport	0.091483
12	StreamingTV	0.013648
13	StreamingMovies	0.010515
14	Contract	0.192666
15	PaperlessBilling	0.006579
16	PaymentMethod	0.017836
17	MonthlyCharges	0.089616
18	TotalCharges	0.099591

Decision Tree



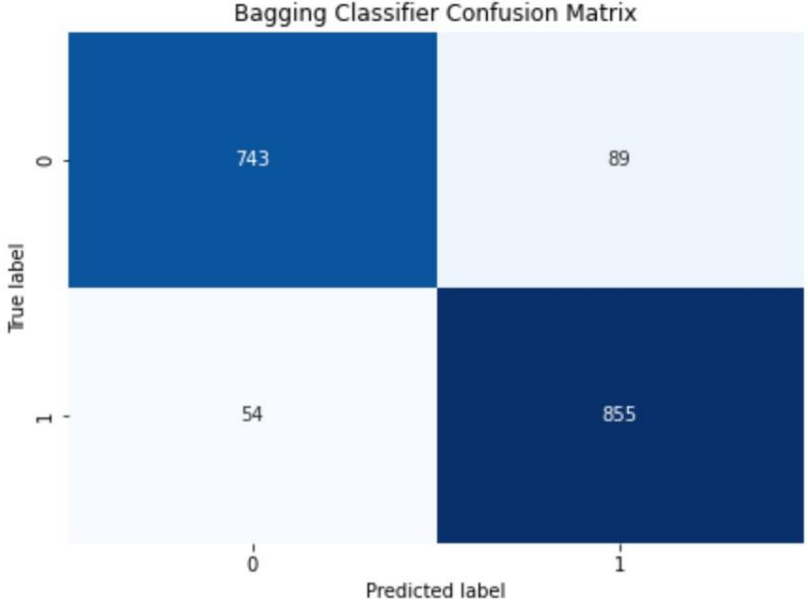

```
Accuracy:  0.9201608271108558
Precision:  0.9157667386609071
Recall:  0.9328932893289329
F1:  0.9242506811989101
Balanced Accuracy:  0.9195716446644664
ROC AUC score:  0.9195716446644664
Cohens Kappa:  0.8398701222219502
```

Logistic Regression

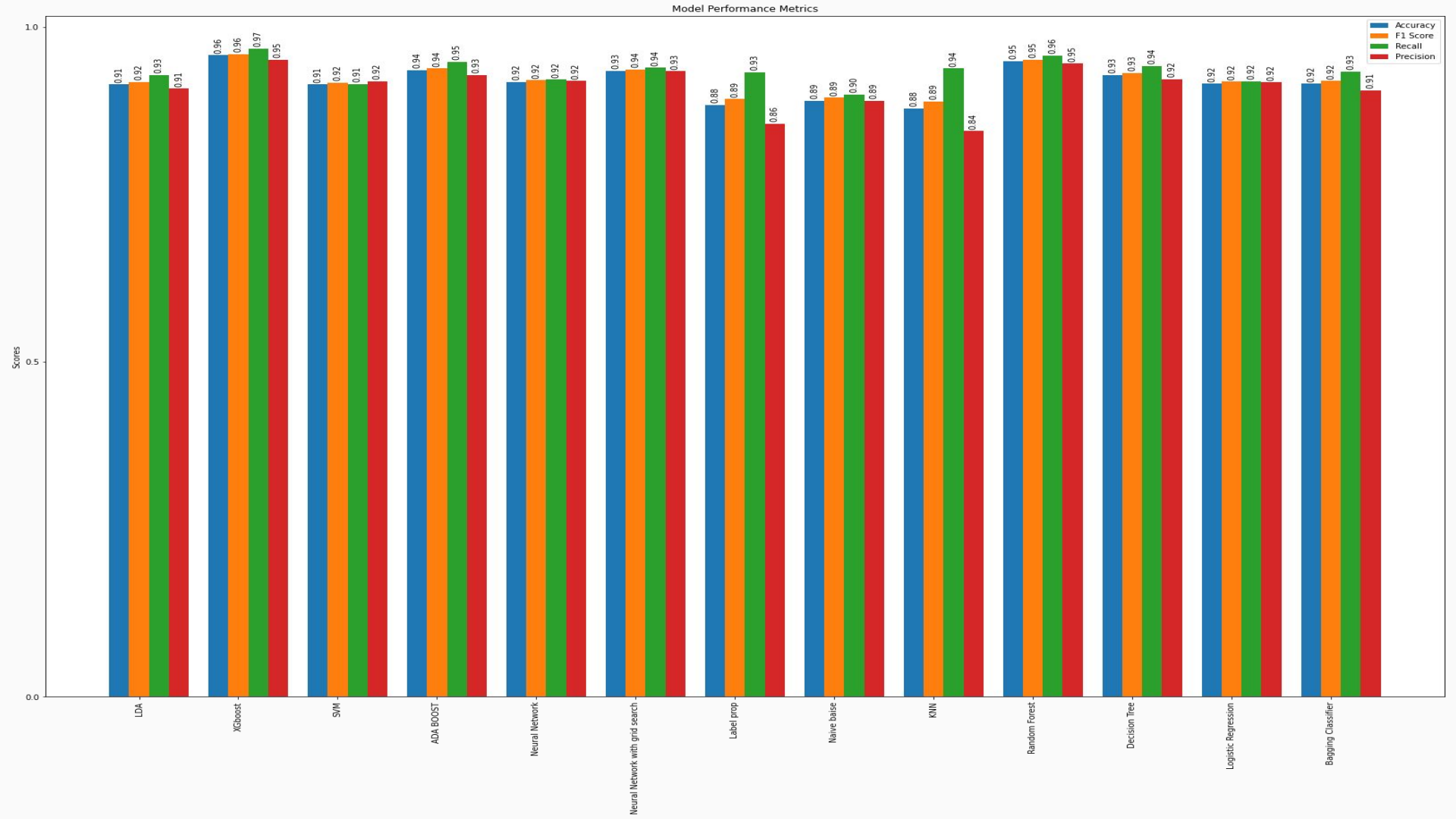



```
Accuracy:  0.9167145318782309
Precision:  0.9161220043572985
Recall:  0.9251925192519251
F1:  0.9206349206349206
Balanced Accuracy:  0.9163222211644242
ROC AUC score:  0.9163222211644241
Cohens Kappa:  0.8330261005646601
```

Bagging Classifier




```
Accuracy:  0.9178632969557725
Precision:  0.9057203389830508
Recall:  0.9405940594059405
F1:  0.9228278467350242
Balanced Accuracy:  0.9168114527798934
ROC AUC score:  0.9168114527798933
Cohens Kappa:  0.8351108464839282
```


Model Performance Metrics
Accuracy
F1 Score
Recall
Precision
Scores
1.0
0.5
0.0
0.91 0.92 0.93 0.91
0.96 0.96 0.97 0.95
0.91 0.92 0.91 0.92
0.94 0.94 0.95 0.93
0.92 0.92 0.92 0.92
0.93 0.94 0.94 0.93
0.88 0.89 0.93 0.86
0.89 0.89 0.90 0.89
0.88 0.89 0.94 0.84
0.95 0.95 0.96 0.95
0.93 0.93 0.94 0.92
0.92 0.92 0.92 0.92
0.92 0.92 0.93 0.91
LDA
XGboost
SVM
ADA BOOST
Neural Network
Neural Network with grid search
Label prop
Naive baise
KNN
Random Forest
Decision Tree
Logistic Regression
Bagging Classifier

Final Conclusion

In conclusion, we have compared various machine learning algorithms for our dataset, including LDA, XGBoost, SVM, AdaBoost, Neural Networks, Label Propagation, Naive Bayes, KNN, Random Forest, Decision Tree, Logistic Regression, and Bagging Classifier. Based on the performance metrics (Accuracy, F1 Score, Recall, and Precision), we observed that XGBoost performed the best among all the models, followed by Random Forest and Neural Network with Grid Search.

We also observe that

- Tenure has the highest importance
- Features such as Partner, Dependents, DeviceProtection, StreamingTV, and StreamingMovies are less important features, with feature importance values ranging from 0.019759 to 0.011085.

Thanks!

Q & A