Identifying Customer Churn

Objective

This project will create a binary classification model to determine which customers are likely to leave the telecom company.

The Dataset

State
Account Length
Area Code
Phone Number
International Plan
Voicemail Plan
Number VM Messages
Total Day Minutes
Total Day Calls
Total Day Charge
Total Eve Minutes
Total Eve Calls
Total Eve Charge
Total Night Minutes
Total Night Calls
Total Night Charge
Total Intl Minutes
Total Intl Calls
Total Intl Charge
Customer Service Calls

Final Feature Set

Account Length
Area Code
International Plan
Voicemail Plan
Number VM Messages
Total Day Minutes
Total Day Calls
Total Eve Minutes
Total Eve Calls
Total Night Minutes
Total Night Calls
Total Intl Minutes
Total Intl Calls
Customer Service Calls

Final Feature Set Importance

total day minutes	0.18	total day calls	0.05
customer service calls	0.14	total night calls	0.05
total eve minutes	0.07	number vm messages	0.03
total intl calls	0.07	area_415	0.03
total night minutes	0.06	intl_yes	0.02
total intl minutes	0.06	vm_yes	0.02
account length	0.05	area_510	0.02
total eve calls	0.05		

Evaluation Criteria

F1 Score – a measure of both:

precision (true positives/predicted positives)

recall (predicted true positives/actual true positives)

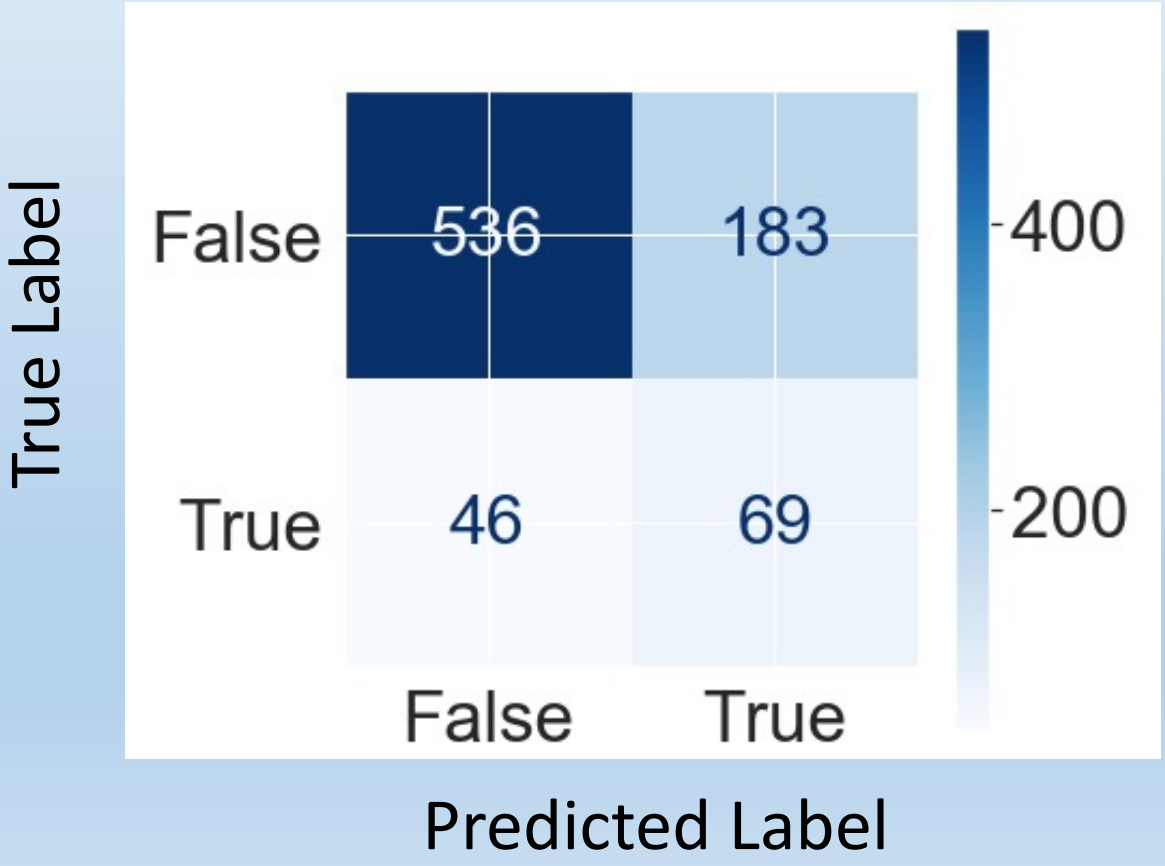
Accuracy –the total number of predictions the model gets correct

Confusion Matrix – shows True Negatives, False Positives

False Negatives, True Positives

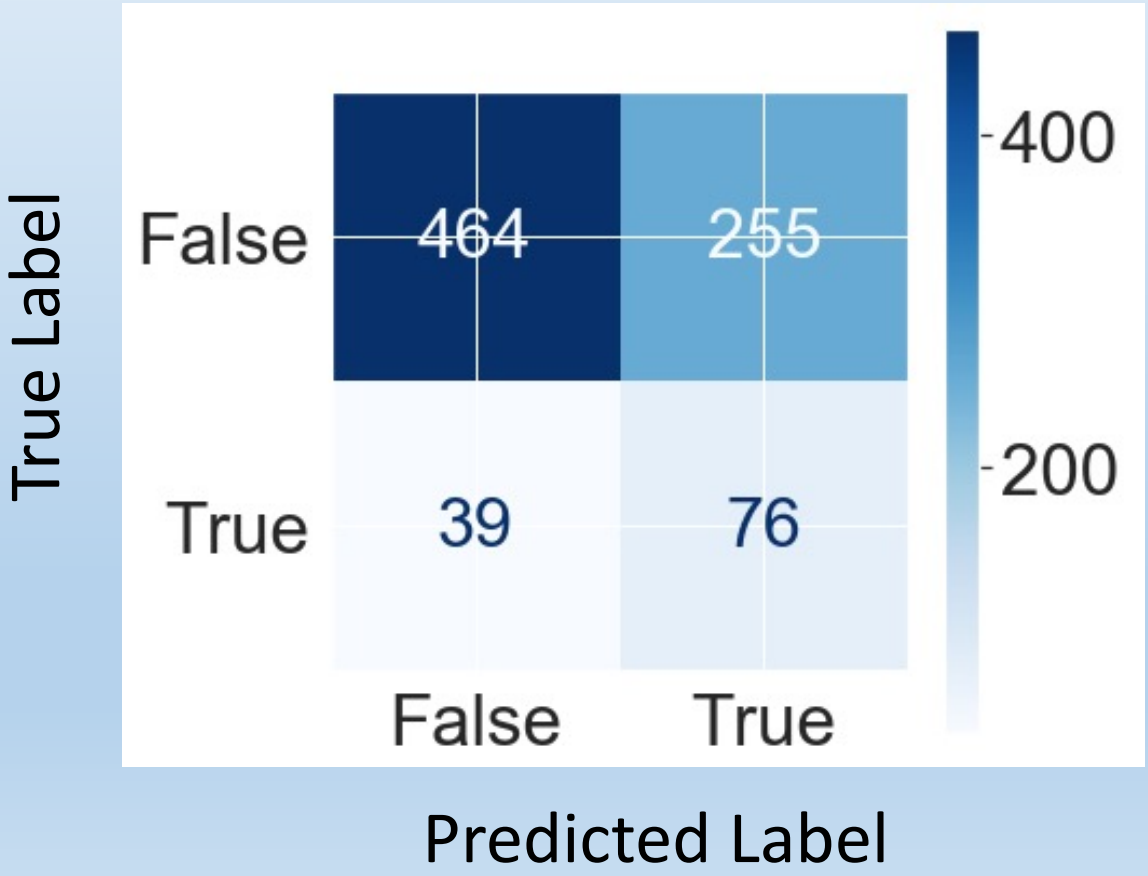
Logistic Regression

F1 Score 0.37 Accuracy 0.71



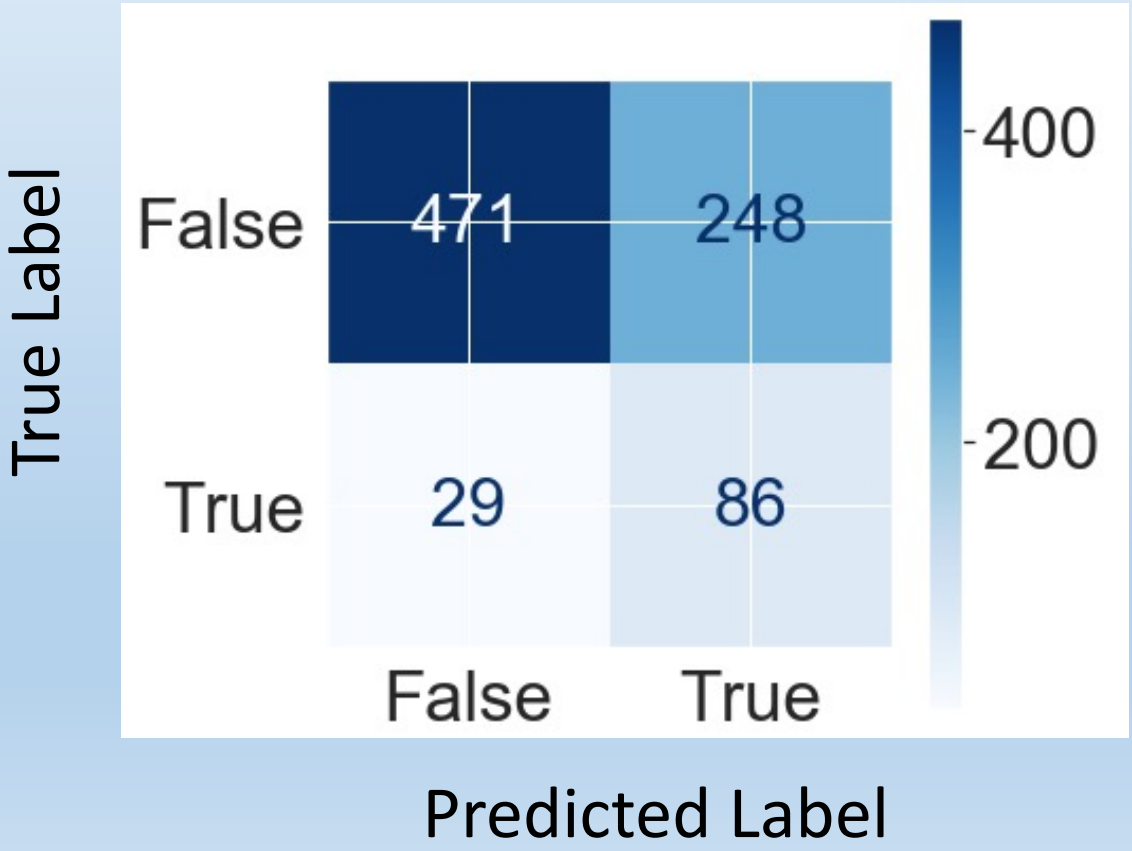
GaussianNB

F1 Score 0.35 Accuracy 0.66



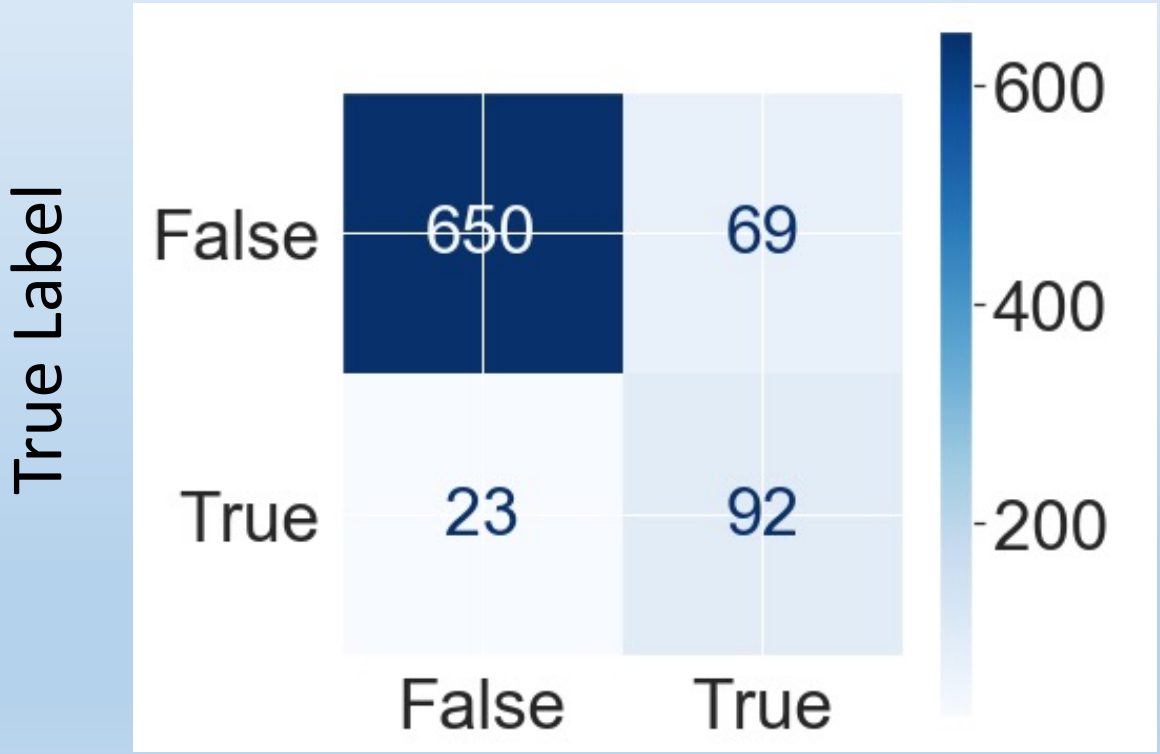
K Nearest Neighbors

F1 Score 0.40 Accuracy 0.70



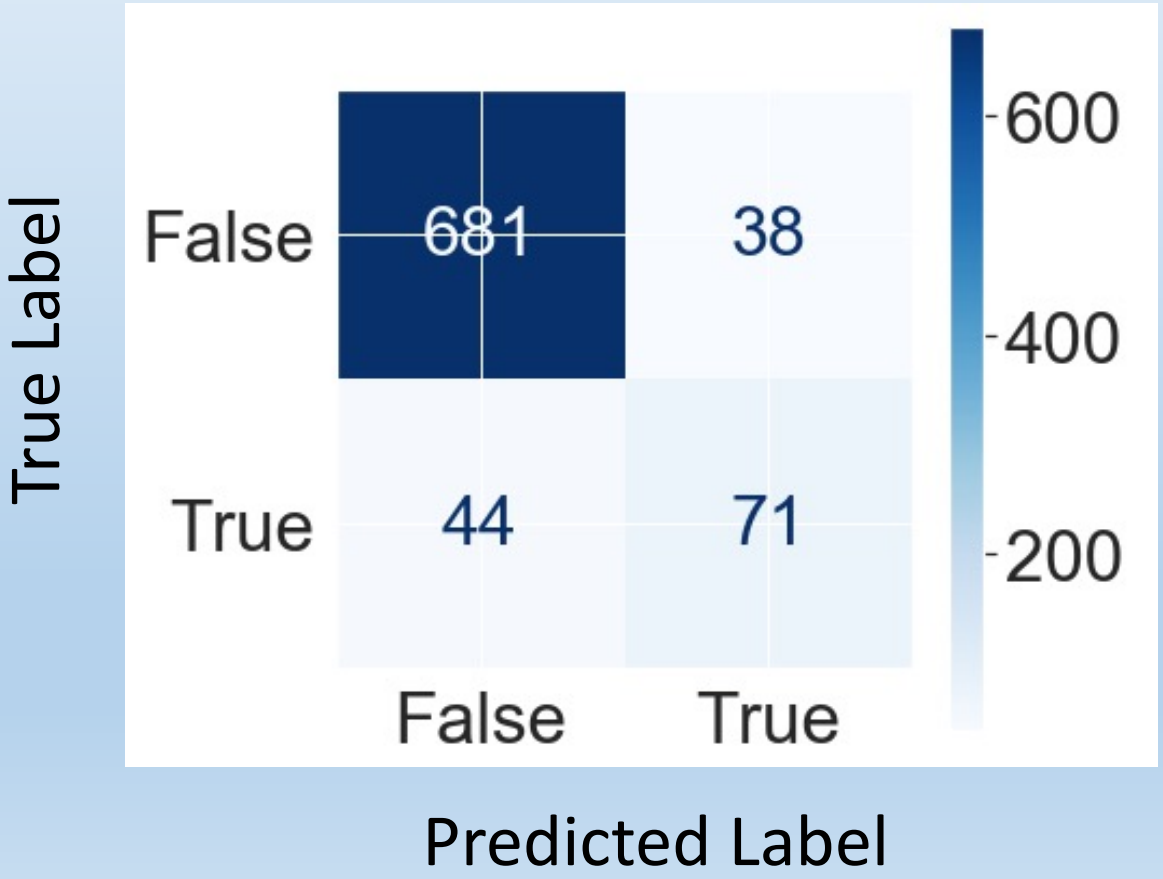
Gradient Boost

F1 Score 0.63 Accuracy 0.88



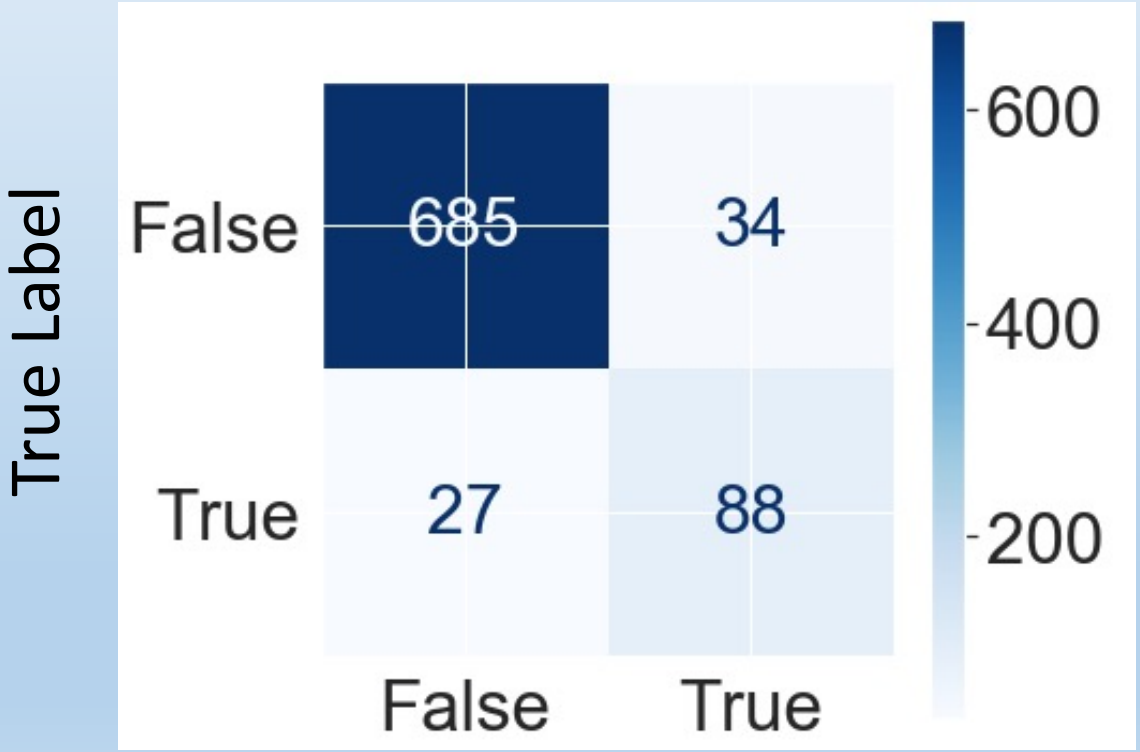
Random Forest

F1 Score 0.61 Accuracy 0.90



XGBoost

F1 Score 0.75 Accuracy 0.93



Conclusion

The XGBoost model gives the best predictive model, with an F1 Score of 0.75 and an accuracy score of 0.93.

The model gives more false positives than negatives, which is preferred as it emphasizes capturing as much churn as possible.

The two most important features for predicting churn are the total number of day minutes used and the number of customer service calls.