

 [kamileyagci](#) / [dsc-phase-3-project](#) Public

forked from [learn-co-curriculum/dsc-phase-3-project](#)

 View license

☆ 0 stars    59 forks

☆ Star

👁 Watch ▾

<> Code

 Pull requests

🎮 Actions

📁 Projects

📖 Wiki

🛡 Security

📈 Insights

 main ▾

...

This branch is 62 commits ahead of [learn-co-curriculum:main](#).

 Contribute ▾

 Fetch upstream ▾



kamileyagci update ...

now 🕒 75

[View code](#)

# SyriaTel Customer Churn Study

Author: Kamile Yagci

Blog URL: <https://kamileyagci.github.io/>

## Overview

In this study, I analyzed the 'SyriaTel Customer Churn' data. The SyriaTel is a telecommunication company. The purpose of the study is to predict whether a customer will ("soon") stop doing business with SyriaTel.

## Business Problem

The telecommincation company, SyriaTel, hired me to analyze the Customer Churn data. The company wants to understand the customer's decision to discontinue their business with SyriaTel. The results of the analysis will be used make business decisions for improving the company finances.

This study will

- Search for the predictable pattern for customer decision on stop or continue doing business with SyriaTel
- Choose a model which will best identify the customers who will stop doing business with SyriaTel

## Method

---

I followed the following steps in this project:

1. Data
  - Load
  - Scrub/Explore
2. Model
  - Pre-Process
  - Evaulation Metrics
  - Logictic Regression
  - K-Nearest Neighbor
  - Decision Trees
  - Random Forest
  - XGBoost
3. Interpret
4. Future work

## Data

---

### Load

I used SyriaTel Customer Churn data for this study. The data file is downloaded from Kaggle.

The file name is 'bigml\_59c28831336c6604c800002a.csv'.

Tha raw data has 3333 entries and 21 columns.

## Scrub/Explore

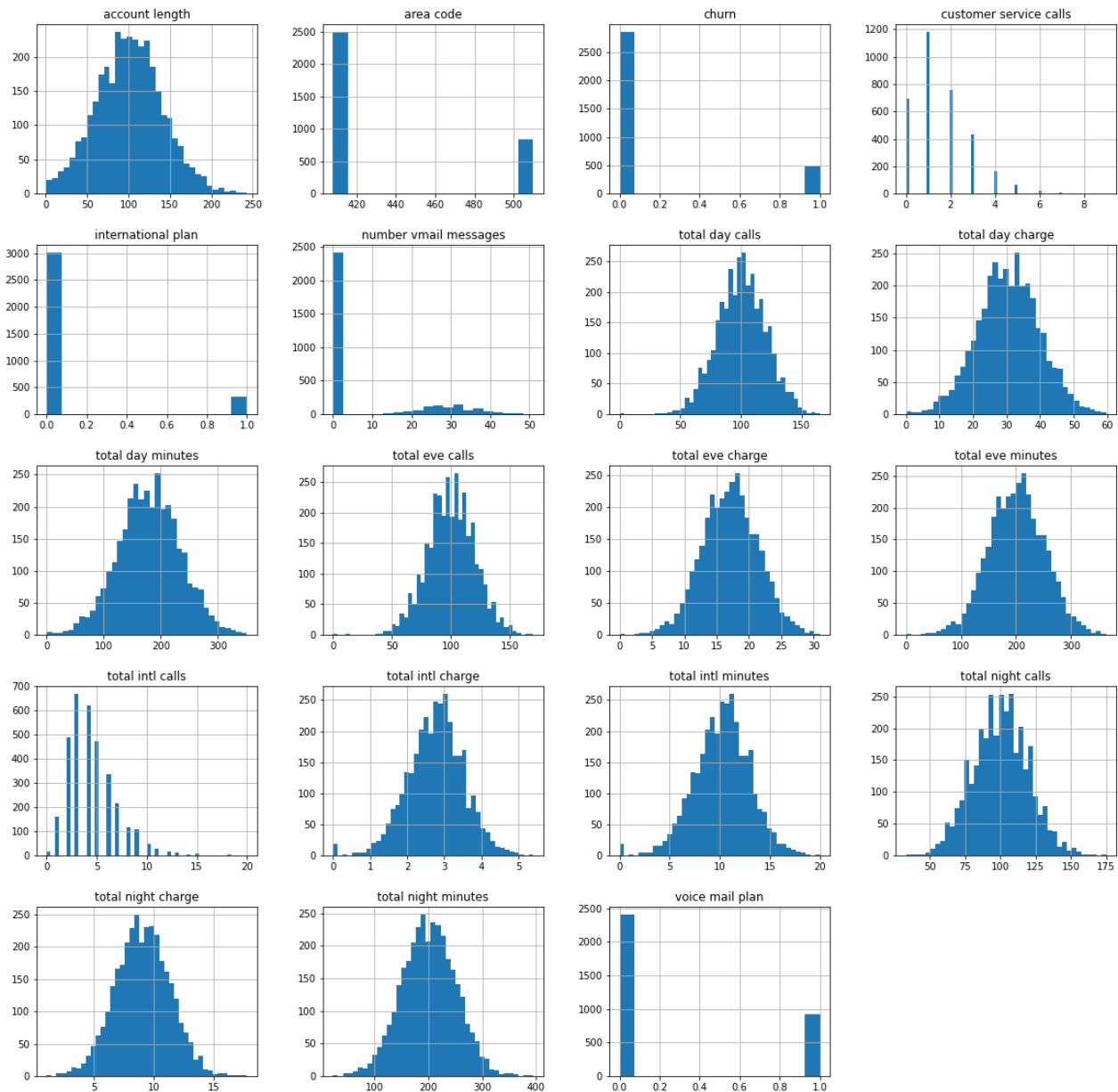
The column/variable names are:

- state
- account length
- area code
- phone number
- international plan
- voice mail plan
- number vmail 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
- churn

The data doesn't have any missing values.

I removed the column 'phone number' from dataset. Most digits in the phone number are random, and it will not have much use in modeling. This variable will also be a problem in dummy variable creation, because each phone number value is unique.

The distribution of variables are shown below. Click on the plot to see them closer.



The target variable for this study is 'churn'. The rest of the variables in the dataset will be predictors.

'churn': activity of customers leaving the company and discarding the services offered

The scatter graphs for 'churn' vs predictors are shown below. Click on the plot to see them closer.



It is hard to recognize any patterns or correlation for 'churn' in these plots.

We will now look at the models to derive patterns and predictions.

## Model

In this study, we are trying to predict customer's decision on stopping the business with the company. The prediction will be True (1) or False (0). Therefore we will use binary classification model.

## Pre-process

Before modeling, I divided the dataset into target data series (y) and predictor dataframe (X).

- y: `DataSeries` of 'churn'
- X: `DataFrame` of all predictors

I also created dummy variables from categorical variables. The X `DataFrame` has 73 variables together with dummies.

Then, I separated the data into train and test splits. I allocated 25% of the data for testing. I also assigned a random state for repeatability.

The shape of the splits:

- X\_train shape = (2499, 73)
- y\_train shape = (2499,)
- X\_test shape = (834, 73)
- y\_test shape = (834,)

shape = (number of rows/entries, number of columns/variables)

The next step is standardization. The data values have different ranges, so I did normalize/scale each variable in train and test data (X) before modeling. I used `Scikit-Learn StandardScaler`.

## Evaluation Metrics

In the next steps, I will use several classifiers to model the data. I will check their performance using the evaluation metrics:

precision:

- Number of True Positives / Number of Predicted Positives
- How precise our predictions are?

recall:

- Number of True Positives / Number of Actual Total Positives

- What percentage of the classes we're interested in were actually captured by the model?

accuracy:

- $(\text{Number of True Positives} + \text{Number of True Negatives}) / (\text{Number of Total Observations})$
- Out of all the predictions our model made, what percentage were correct?

f1-score:

- $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$
- Harmonic Mean of Precision and Recall.

*Source: Flatiron Data Science Curriculum, Evaluation Metrics*

Since my business problem is focusing on identifying the customers who stop doing business, I am interested mainly on the 'recall' metrics. However, when optimizing my model, I should also pay attention to the 'precision'. I want my predictions to be true, to be precise. The recall and precision are inversely proportional. Therefore, I choose to use the f1-score, Harmonic Mean of Precision and Recall, as the main metric for evaluating the performance of the model.

## Logistic Regression

I started modeling with Logistic Regression classifier (LogisticRegression). I instantiated the model with default parameters and fit on training data. Then I checked the evaluation metrics both for training and testing data.

	f1-score	recall	precision	accuracy
Train	0.37	0.27	0.64	0.85
Test	0.32	0.22	0.56	0.86

- The metrics look similar for both training and testing data, just training is a bit better; so slight overfitting.
- The precision - recall - f1 scores are low (for churn=1), so the model prediction performance is not good.
- The accuracy score is high, but misleading. It is caused by the imbalanced dataset.

## Class Imbalance

Class imbalance effects the performance of the classification model. I have looked at the class distributions for the whole data: train + test:

	Value Counts in whole dataset	Normalized
churn = 0	2850	0.855
churn = 1	483	0.145

According to the dataset, 85.5% of the customers do continue with SyriaTel and 14.5% of customers stop business. If we make a prediction saying all customers will continue business, then we will have about 85.5% accuracy. This explains the high accuracy score of the model, despite the other low metric values.

I used SMOTE to create a synthetic training sample to take care of imbalance. After the resampling, the value counts in each class, in training data sample, became equal.

	Original training data, Value counts	Synthetic training data, Value counts
churn = 0	2141	2141
churn = 1	358	2141

I have then reapplied the Logistic Regression, using the resampled training data. The results:

	f1-score	recall	precision	accuracy
Train	0.80	0.81	0.79	0.80
Test	0.51	0.77	0.38	0.78

- After resampling, the Logistic Regression Model performance is clearly improved.
- The performance in training data is better than test data. This is a sign of overfitting.

I initially used the default parameters for the Logistic Regression model. I then applied parameter tuning with GridSearchCV. It determined the best parameter combination for the given parameter grid. I used the f1-score for tuning.

**The results of parameter tuning:**

Best Parameter Combination: {'C': 0.01, 'solver': 'liblinear'}



	<b>f1-score</b>	<b>recall</b>	<b>precision</b>	<b>accuracy</b>
Train	0.80	.83	0.78	0.79
Test	0.51	0.79	0.37	0.77

- It looks like the parameter tuning, with the given parameter grid, didn't improve the performance much.
- Overfitting is observed.

## K-Nearest Neighbors

My next classifier is K-Nearest Neighbors (KNeighborsClassifier). I used the resampled training data for fitting the model with default parameters.

	<b>f1-score</b>	<b>recall</b>	<b>precision</b>	<b>accuracy</b>
Train	0.92	1.00	0.85	0.91
Test	0.39	0.62	0.29	0.72

- The fitting on resampled training data has a better performance. The f1-score for test data increased from 0.15 to 0.39. (The results for resampled data is not shown here, but tested).
- Overfitting observed.

Then, I used GridSearchCV for parameter tuning.

### The results of parameter tuning:

Best Parameter Combination: {'n\_neighbors': 4, 'p': 1}

	<b>f1-score</b>	<b>recall</b>	<b>precision</b>	<b>accuracy</b>
Train	0.97	0.99	0.94	0.97
Test	0.35	0.39	0.32	0.79

- Parameter tuning, with the given parameter ranges, didn't improve the KNN model performance.
- Overfitting observed.

## Decision Tress

I firstly used DecisionTreeClassifier with default parameters, then applied GridSearchCV to find the optimum parameteres.

	f1-score	recall	precision	accuracy
Train	1.00	1.00	1.00	1.00
Test	0.64	0.67	0.62	0.89

### The results of parameter tuning:

Best Parameter Combination: {'criterion': 'gini', 'max\_depth': 6, 'min\_samples\_split': 2}

	f1-score	recall	precision	accuracy
Train	0.90	0.87	0.94	0.91
Test	0.69	0.70	0.68	0.91

- The parameter tuning improved the Decision Trees performance a little.
- Overfitting observed.

## Random Forests

Next, I used ensemble method Random Forests (RandomForestClassifier), which uses DecisionTreeClassifier.

	f1-score	recall	precision	accuracy
Train	1.00	1.00	1.00	1.00
Test	0.75	0.71	0.79	0.93

### \*The results of parameter tuning:

Best Parameter Combination: {'criterion': 'gini', 'max\_depth': 6, 'max\_features': 8, 'min\_samples\_split': 6, 'n\_estimators': 10}

	f1-score	recall	precision	accuracy
Train	0.90	0.89	0.92	0.91
Test	0.65	0.69	0.61	0.89

- The paremeter tuning didn't improve the performance of Random Forest model.

- Overfitting observed.

## XGBoost

Last, I used another ensemble method XGBoost (XGBClassifier).

	f1-score	recall	precision	accuracy
Train	1.00	1.00	1.00	1.00
Test	0.81	0.73	0.90	0.95

### The results of parameter tuning:

Best Parameter Combination: {'learning\_rate': 0.1, 'max\_depth': 6, 'min\_child\_weight': 1, 'n\_estimators': 100, 'subsample': 0.7}

	f1-score	recall	precision	accuracy
Train	0.99	0.98	1.00	0.99
Test	0.81	0.75	0.89	0.95

- The parameter tuning didn't effect the XGBoost performance much.
- Overfitting observed.

## Compare the models

I compared the classification models to choose the best one that identifies the customers who will study doing business with SyriaTel .

I looked at the evaluation metrics like precision, recall, accuracy and f1.

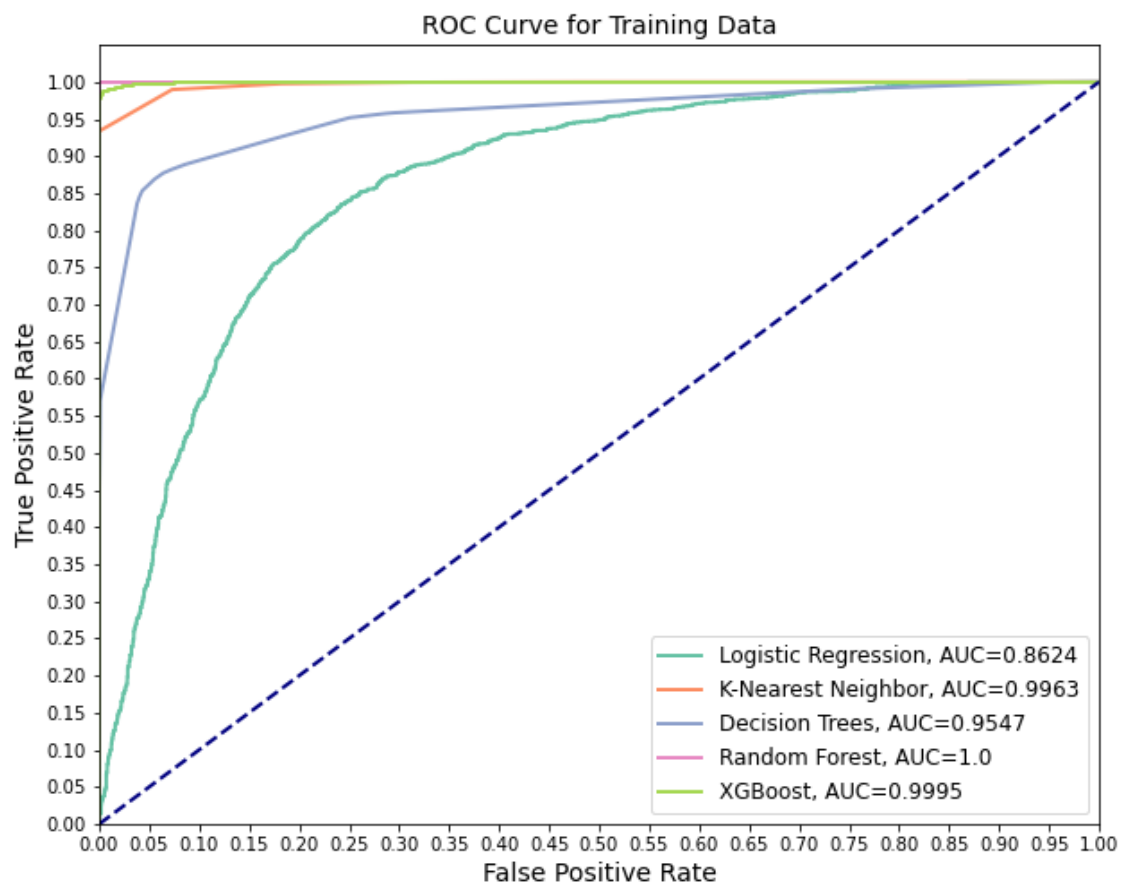
I also plotted ROC curves and calculated AUC for each model.

- ROC: Receiver Operating Characteristic curve illustrates the true positive rate against the false positive rate.
- AUC: Area Under Curve

I used the optimal/best parameter set to instantiate my models. For some models, the GridSearchCV selected the parameters which causes large overfitting; so low performance on test data. I used Default parameters for these models.

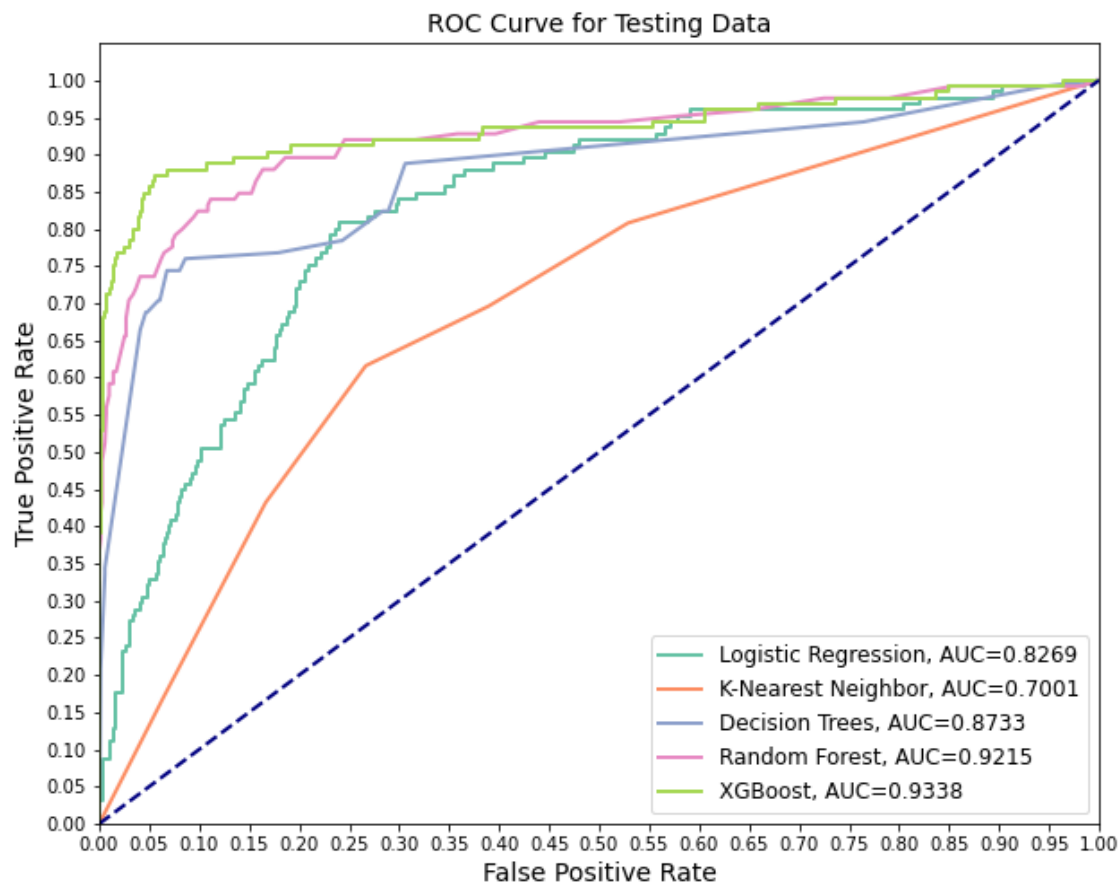
### The evaluation metrics and ROC Curve for Training data:

	precision	recall	accuracy	f1	auc
model					
<b>Logistic Regression</b>	0.777	0.825	0.794	0.801	0.862
<b>K-Nearest Neighbor</b>	0.845	0.998	0.908	0.915	0.996
<b>Decision Trees</b>	0.940	0.869	0.907	0.903	0.955
<b>Random Forest</b>	1.000	1.000	1.000	1.000	1.000
<b>XGBoost</b>	0.998	0.984	0.991	0.991	1.000



The evaluation metrics and ROC Curve for Test data:

	precision	recall	accuracy	f1	auc
model					
<b>Logistic Regression</b>	0.372	0.792	0.769	0.506	0.827
<b>K-Nearest Neighbor</b>	0.289	0.616	0.716	0.394	0.700
<b>Decision Trees</b>	0.682	0.704	0.906	0.693	0.875
<b>Random Forest</b>	0.788	0.712	0.928	0.748	0.922
<b>XGBoost</b>	0.887	0.752	0.948	0.814	0.934



All of my models showed some pattern for customer decision on stop or continue doing business. They also did predictions to identify the customers who will discontinue service (churn customers).

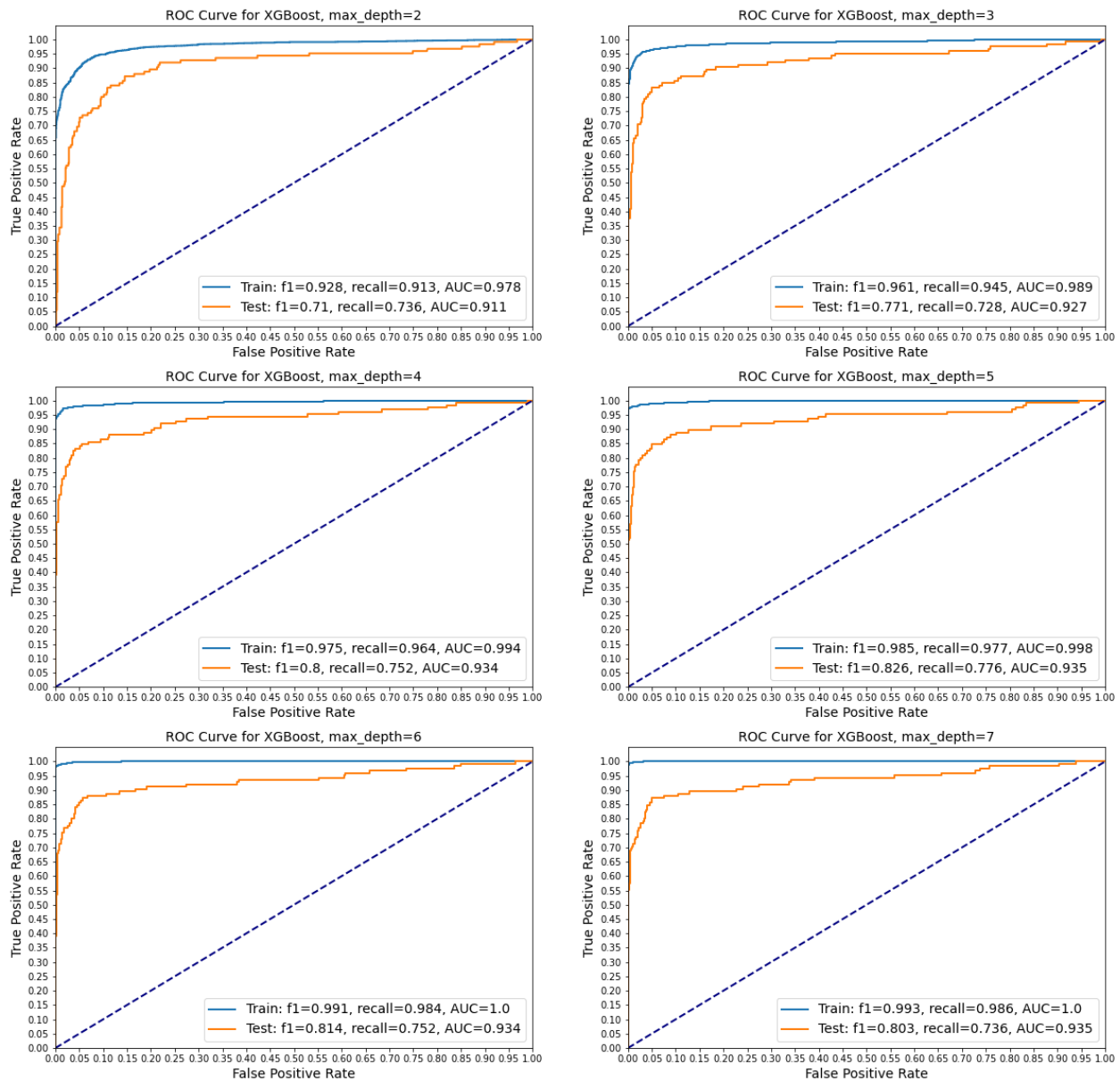
Which model is best on identifying churn customers?

According to the test data evaluation metrics, the XGBoost classifier has overall best performance. It also has the best 'recall' and 'f1 score', which matters most for my study.

I choose the XGBoost Classifier as the best model.

## Overfitting in XGBoost model

The XGBoost model performed better in training data than the test data. This is overfitting. The decreasing the 'max\_depth' can help to minimize the overfitting. I plotted ROC curve for several max\_depth values to observe the overfitting.



The overfitting decreased a little bit, when max\_depth is 4 or 5. The performance of the

## Final Model

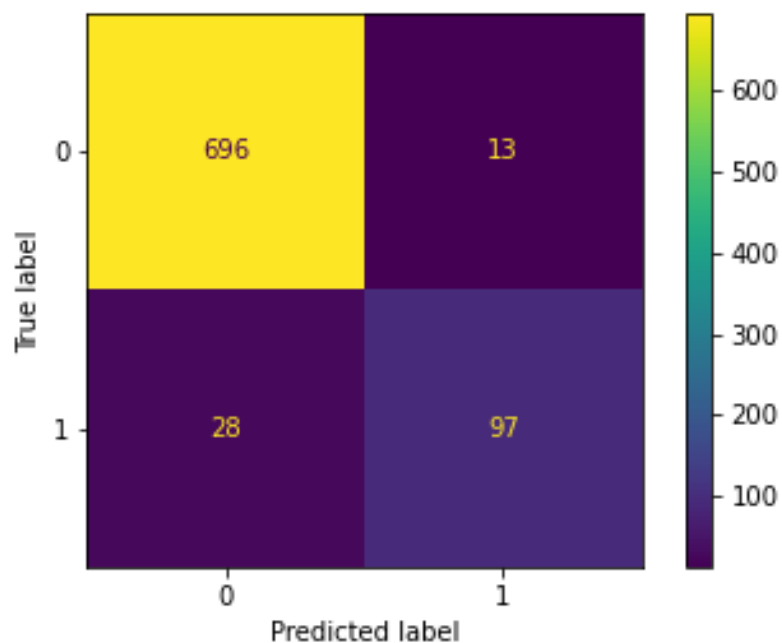
I create my final model with XGBoost Classifier with the below parameters.

```
{'learning_rate': 0.1, 'max_depth': 5, 'min_child_weight': 1, 'n_estimators': 100, 'subsample': 0.7}
```

The evaluation metrics for final, XGBoost model:

	f1-score	recall	precision	accuracy
Train	0.98	0.98	0.98	0.99
Test	0.83	0.78	0.88	0.95

The confusion matrix for final model:



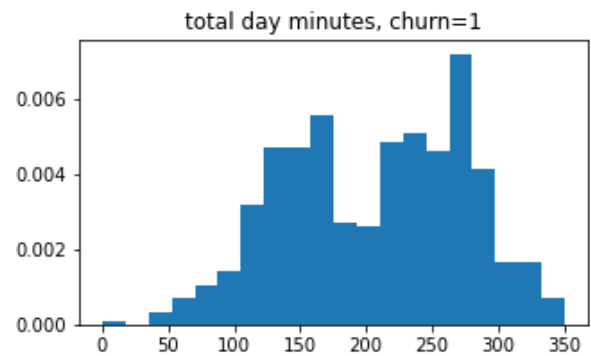
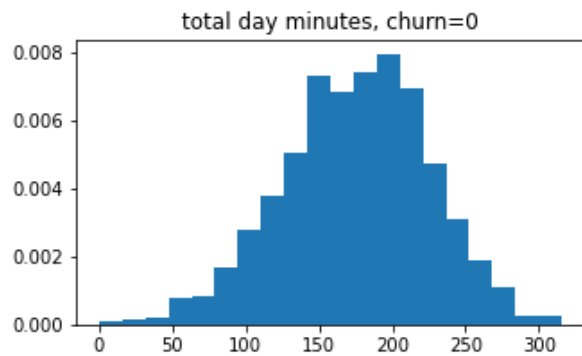
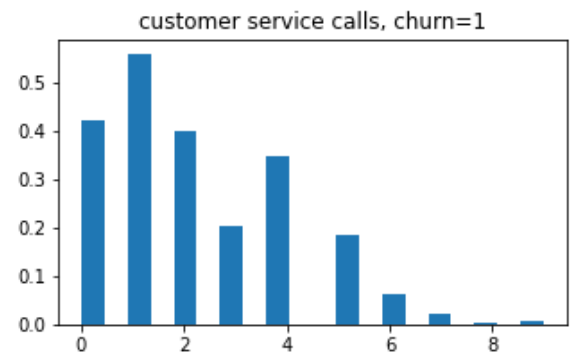
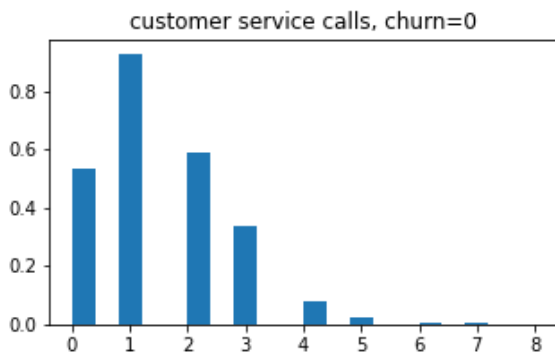
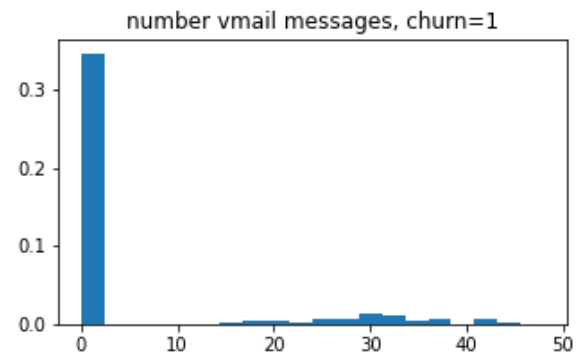
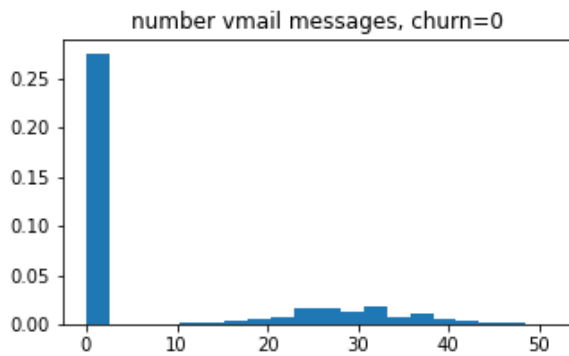
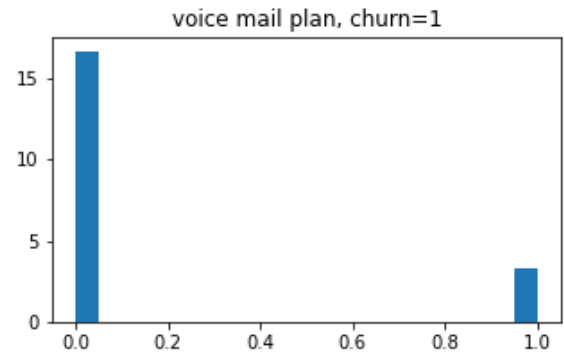
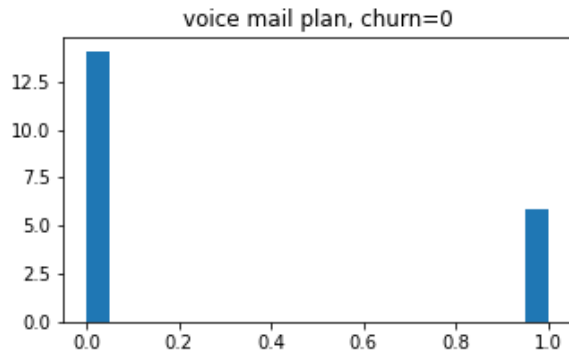
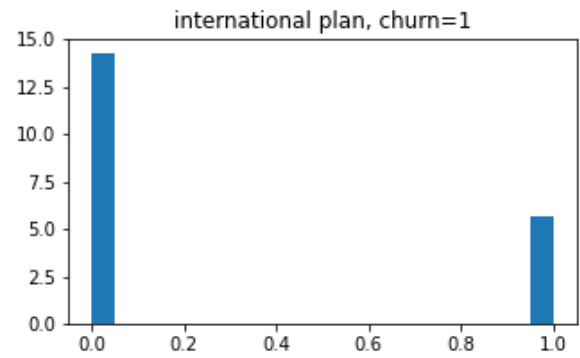
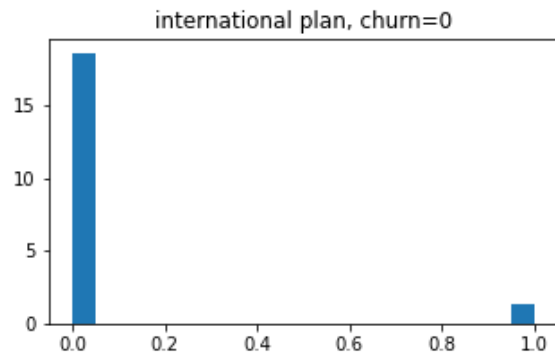
Important Features for the final model:

The top 10 important features according to the final model:

	<b>predictors</b>	<b>importance</b>
<b>69</b>	international plan_0	0.099607
<b>71</b>	voice mail plan_0	0.090104
<b>14</b>	customer service calls	0.076601
<b>1</b>	number vmail messages	0.056749
<b>2</b>	total day minutes	0.044972
<b>67</b>	area code_415	0.034792
<b>7</b>	total eve charge	0.033760
<b>68</b>	area code_510	0.032788
<b>41</b>	state_MT	0.024700
<b>66</b>	area code_408	0.024159

For churn=0 and churn=1 data, the distribution of top 5 important features are plotted:





# Interpret

---

The summary of Final Model performance:

- It successfully identifies the 78% of the true churn customers. (recall)
- Among the model predicted churn customers, 88% of them are true churn customers. (precision)
- The Harmonic Mean of Precision and Recall (f1-score) is 83%.

The identification numbers on test data:

- Identification numbers:
  - Number of true positives: 97
  - Number of true negatives: 696
  - Number of false positives: 13
  - Number of false negatives: 28
- It identifies 97 out of 125 churn customers correctly.
- 97 out of 110 predicted churn customers are real churn.

Characteristic of churn customers:

- The churn customers are more likely to have international plan than continuous customers.
- The churn customers are less likely to have voice mail plan than continuous customers.
- The churn customers have less voice mail messages than continuous customers (as a result of less voice mail plan)
- The churn customers have more customer service calls than continuous customers.
- The churn customers have more total day minutes than continuous customers.

## Business Recommendations

---

- Improve the international plan to keep/attract customers
- Improve customer service center for better customer satisfaction
- Offer a discounted plan with high total day minutes

## Future Work

---

- Improve the XGBT model (final model) performance

- Search each parameter separately to understand the effect on performance
- Obtain a more sensitive/informed range for each parameter to be used in grid search
- Study the effect of other hyperparameters
- Use weighted f1-score, with more weight on recall than precision
  - to compare model performance
  - and for parameter tuning

## Releases

No releases published

[Create a new release](#)

---

## Packages

No packages published

[Publish your first package](#)

---

## Languages

● **Jupyter Notebook** 100.0%