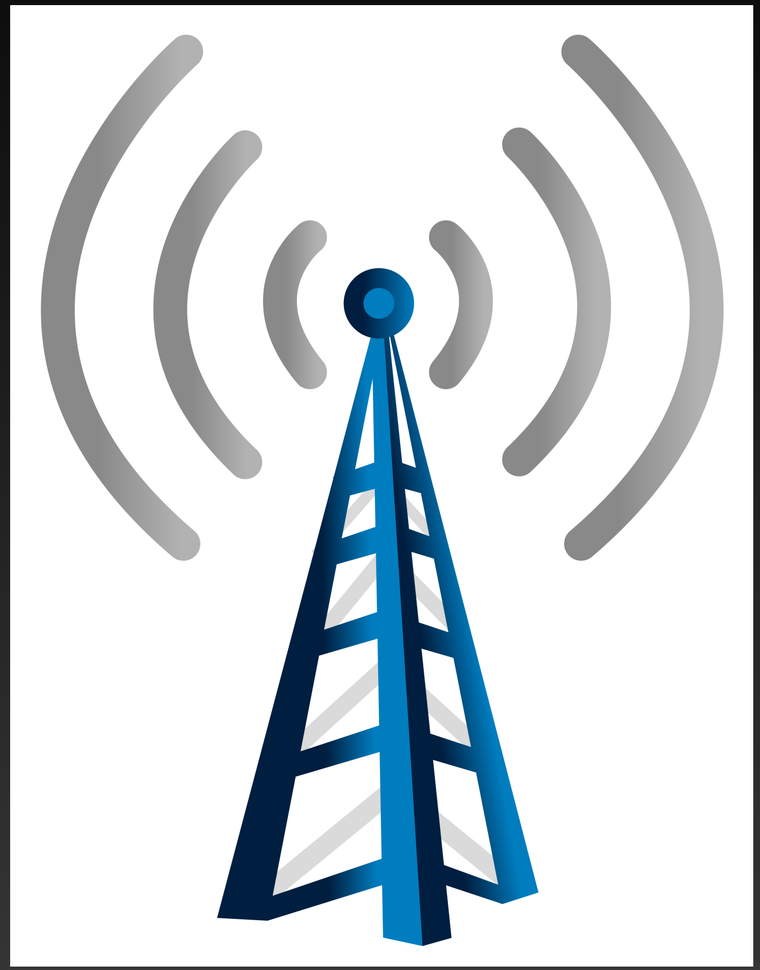
SyriaTel Customer Churn Study

Flatiron School

Kamile Yagci 1/07/2021

Content

- Overview
- Business Problem
- Data
- Method
- Model
- Interpret
- Future Work
- Questions



Overview

- This project is the analysis of '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.
- The Data Classification models are used in analysis.

Project Details:

- GitHub: https://github.com/kamileyagci/dsc-phase-3-project

Business Problem

The telecommunication 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 improve 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

Data

- SyriaTel Customer Churn data: ‘bigml_59c28831336c6604c800002a.csv' (source Kaggle)
- The data contains 3333 entries/customers.
- For each customer, the data includes 21 types of information:
  - state, area code, phone number
  - international plan, total intl minutes, total intl calls, total intl charge
  - voice mail plan , number vmail messages
  - total day minutes, total day calls, total day charge
  - total evening minutes, total evening calls, total evening charge
  - total night minutes, total night calls, total night charge
  - customer service calls
  - account length
  - churn (= activity of customers leaving the company and discarding the services offered)

Model (1)

- Goal: Predict ‘churn’ value
  - churn: activity of customers leaving the company and discarding the services offered
  - The prediction will be True (1) or False (0)
  - Binary classification is used for modeling.
- Pre-processing:
  - Divided the dataset into y: target ‘churn’ and X: all predictors
  - Standardize the data
  - Split data into training and testing subsets:
    - Training data: used to train our model
    - Testing data: used to validate the model

Model (2)

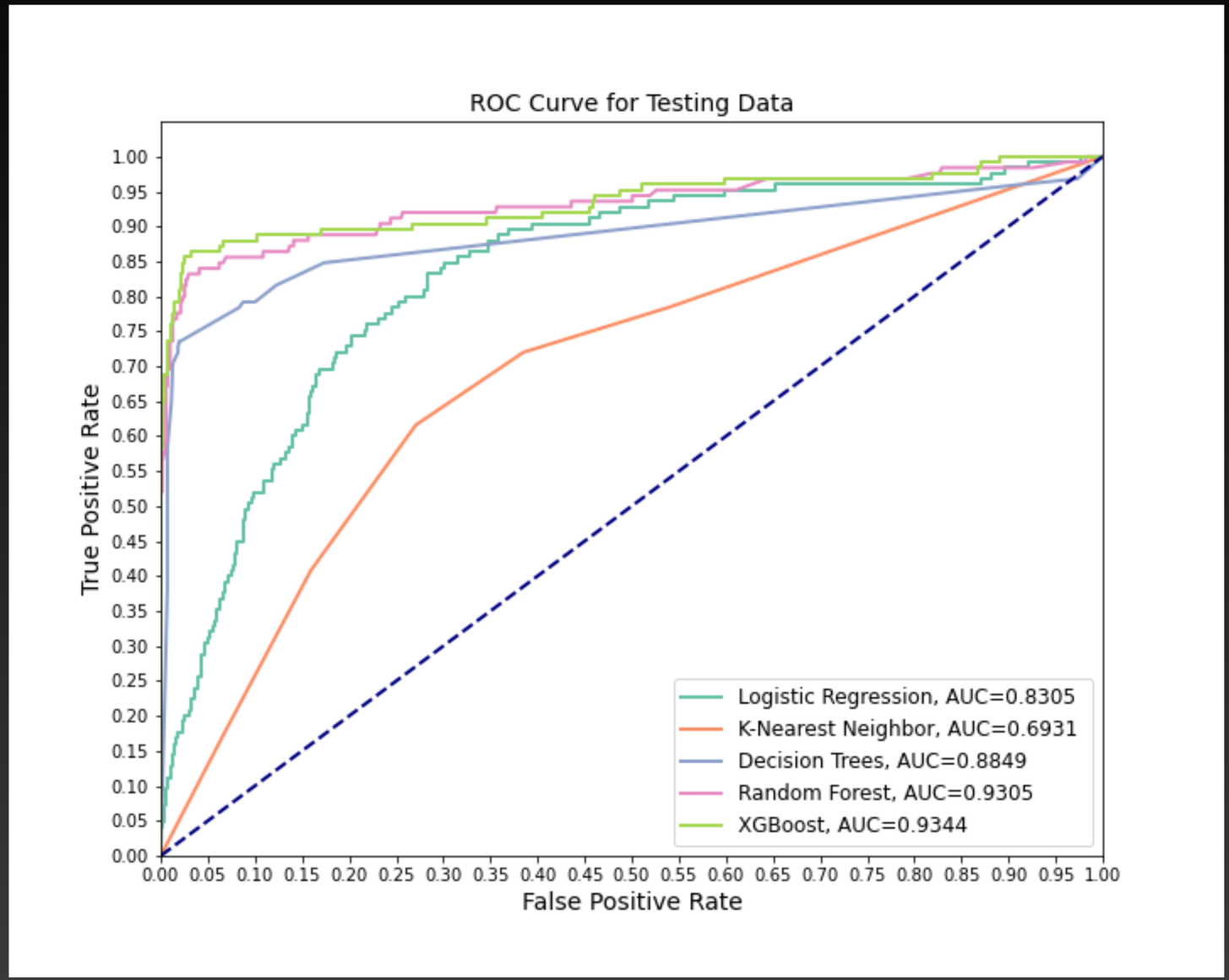
- Several classifier models are tested:
  - Logistic Regression
  - K-Nearest Neighbors
  - Decision Trees
  - Random Forest
  - XGBoost
- Each model is optimized based on f1-score

Model (3)

- Evaluation metrics are used to measure the performance of the models:
  - precision: #of True Positives / # of Predicted Positives
    - What percentage of model predictions are true?
  - recall: # of True Positives / # of Actual Total Positives
    - What percentage of the classes we're interested in were actually captured by the model?
  - accuracy: (# of True Positives + # of True Negatives) / (# of Total Observations)
    - Out of all the predictions our model made, what percentage were correct?
  - f1-score: 2 * (Precision * Recall) / (Precision + Recall)
    - Harmonic Mean of Precision and Recall
- I focus on ‘f1-score’ and ‘recall’ for model comparison based on the goal

Model (4)

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



Model (5)

- Evaluation Metric Scores:

model	precision	recall	accuracy	f1	auc
Logistic Regression	0.330189	0.840	0.720624	0.474041	0.830511
K-Nearest Neighbor	0.286245	0.616	0.712230	0.390863	0.693106
Decision Trees	0.873786	0.720	0.942446	0.789474	0.884897
Random Forest	0.961039	0.592	0.935252	0.732673	0.930482
XGBoost	0.948454	0.736	0.954436	0.828829	0.934409

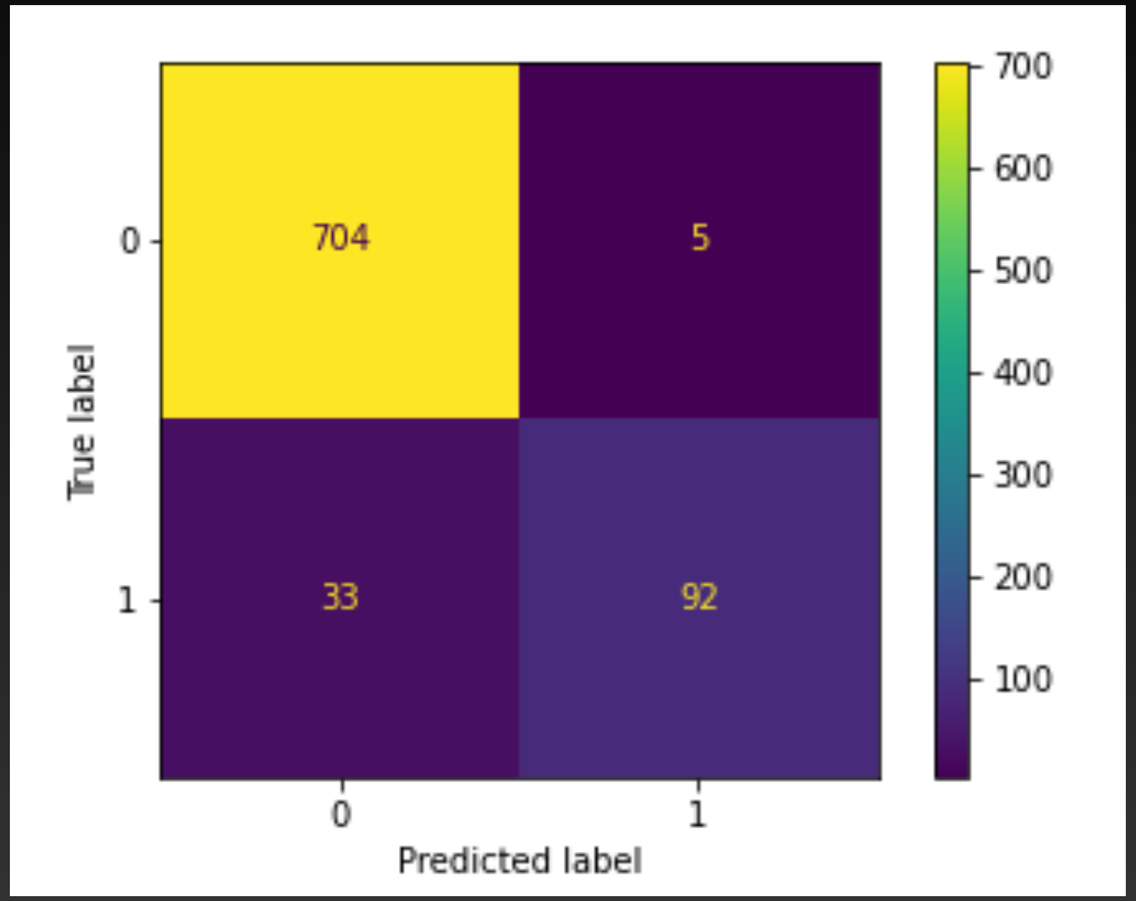
Interpret (1)

- Final Model = XGBoost Classifier Model
  - best 'recall' and ‘f1-score’
- Final Model identifies the 74% of the true churn customers. (recall)
- Among the model predicted churn customers, 95% of them are true churn customers. (precision)
- The Harmonic Mean of Precision and Recall (f1-score) is 83%.

churn = activity of customers leaving the company and discarding the services offered

Interpret (2)

- The confusion matrix of test data for final model
  - Number of true positives: 92
  - Number of true negatives: 704
  - Number of false positives: 5
  - Number of false negatives: 33
- Identifies 92 out of 125 churn customers correctly (74% recall).
- 92 out of 97 predicted churn customers are real churn (95% precision).



Future Work

- Improve the XGBoost model (final model) performance
  - Search each parameter separately to understand the effect on performance
  - Obtain a more sensitive/informed parameter range 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

Questions?