Telecom Churn Prediction 2020

Calvin Yu

Problem/Need

A telecom company puts thousands of money on a direct marketing campaign to keep its customers loyal.

- Like to know what factors cause customers to churn
- Want to spot potential churn
- Want to know what to promote on to prevent customer churn

Impact Hypothesis

- Create a classification algorithm can predict who has a high probability to churn and understand what features are contributing to churn rate the most
- Spot potential churn customer
- Know what to focus on when promoting a marketing campaign

Solution Path

1. Exploratory Data Analysis : Create diagrams to understand the dataset
2. Pick a classification metric - F1 score and Accuracy
3. Test different algorithms
4. Pick the best algorithm with the highest score on the metrics

Scoping Considerations

- **Assumption:**

  People who churn share similar features
- **Measure of Success**

  F1 score

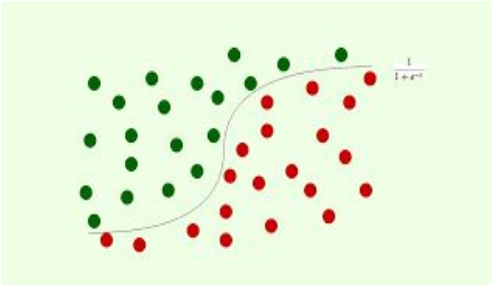
  Accuracy score
- **Risk/Difficulty**

  Some algorithms are hard to interpret

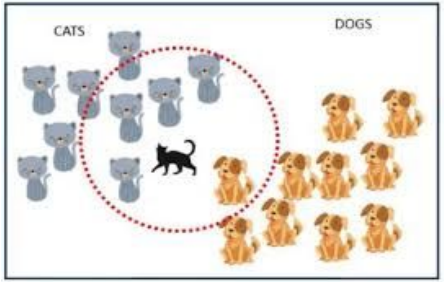
Data

- Customer Churn Prediction 2020 from Kaggle
- Training : 4250 rows with 20 columns with one of them is the target variable
- Testing: 750 rows with same columns but not including the target variable

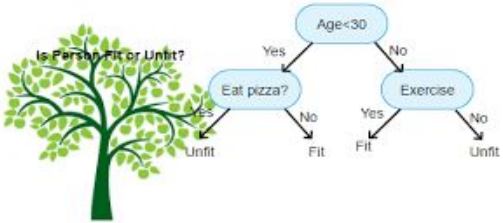
Algorithms



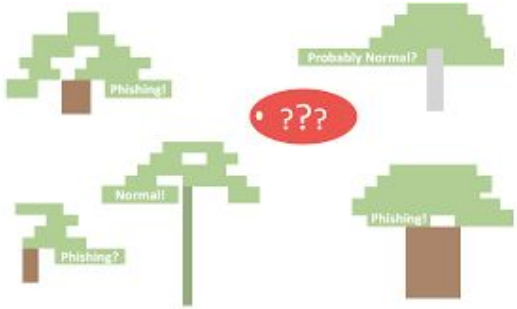
Logistic Regression



K-Nearest Neighbor

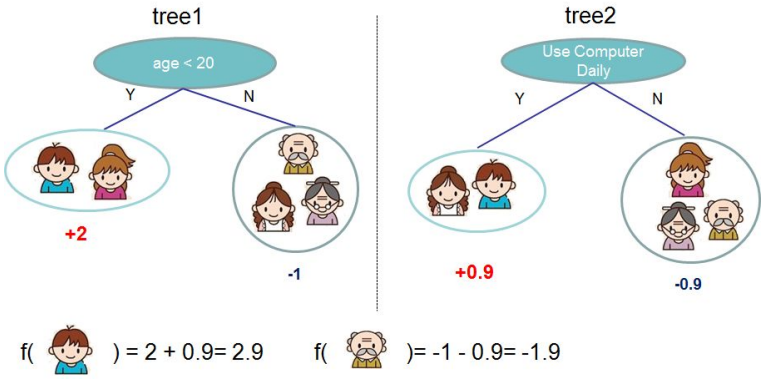


Decision Tree



Random Forest

Algorithms con.



Gradient Boost / XGBoost

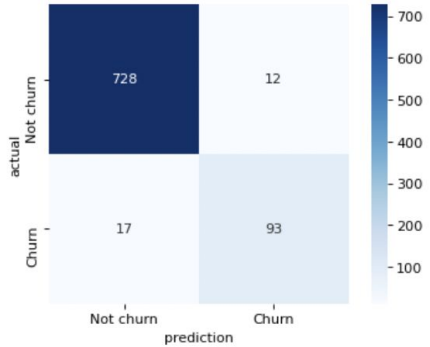
Naive Bayes

Results on validation set

```
Recall score 0.6205797101449275
Precision score 0.8721121098342625
f1 score 0.7245780729893584
roc_auc score 0.9048642338797619
accuracy score 0.935686274509804
              precision    recall  f1-score   support

       False       0.98      0.98      0.98       740
        True       0.89      0.85      0.87       110

    accuracy                           0.97       850
   macro avg       0.93      0.91      0.92       850
weighted avg       0.97      0.97      0.97       850
```



Best model is XGboost

~ 0.73 F1 score

~ 0.94 Accuracy Score

Results on actual test set

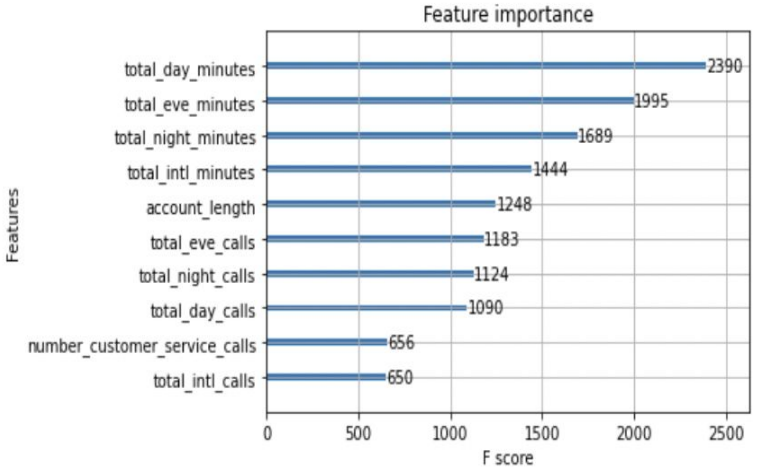
Private Score	Public Score
0.93523	0.97777

Accuracy score

~ 0.98

Feature Importances

```
<AxesSubplot:title={'center':'Feature importance'}, xlabel='F score', ylabel='Features'>
```



- Total minutes per daytime,evening, and nighttime are the top 3 important
- Account length is the 4th important

Communication

- The total minutes per day despite the time is important, so the marketing campaign can be done with giving out discounts if a customer exceeds certain minutes per month
- Account length is also important, so the telecom company can give reward to customers who have been subscribed for a certain of time

Further step

- Create marketing campaign with the slogan of giving out discount on minutes per month to attract new customers
- Reward the customers who have been subscribed for a long period of time
- Coldcall the customers who have churned to understand the reason of them churing