Syriatel's Customer Retention Analysis

By: Zach Hyde

Overview

This particular dataset intrigued be because I have been in the customer service industry and currently work within retail. My aspirations are to become an effective Data Scientist where I can utilize my skill-set to bring reliable insight to my business partner(s) where they can take my analysys into high consideration when making decisions to strengthen the whole operation. The dataset that I have analyzed for you today represents information gathered by Telecom of their customers utilization of their product in a variety of features while also describing if that customer churned or not. The goal for this project was to come up with a model that effectively utilized certain features to accurately predict whether or not the customer would continue services or cancel them.

Data Usage

- [Kaggle sourced CSV]

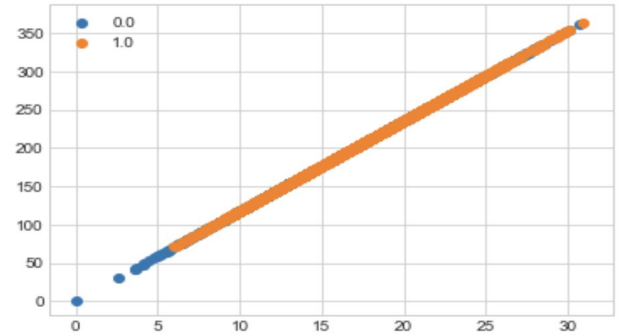
  - Overview of Data
    - Features - 20
    - Data points - 3333
    - Target class - ‘Churn’
      - Imbalanced ratio ~ 85/15% (no/yes)

```
<class 'pandas.core.frame.DataFrame'>
Index: 3333 entries, KS to TN
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   account length          3333 non-null   int64
 1   area code               3333 non-null   int64
 2   phone number            3333 non-null   object
 3   international plan      3333 non-null   object
 4   voice mail plan         3333 non-null   object
 5   number vmail messages   3333 non-null   int64
 6   total day minutes       3333 non-null   float64
 7   total day calls         3333 non-null   int64
 8   total day charge        3333 non-null   float64
 9   total eve minutes       3333 non-null   float64
 10  total eve calls         3333 non-null   int64
 11  total eve charge        3333 non-null   float64
 12  total night minutes     3333 non-null   float64
 13  total night calls       3333 non-null   int64
 14  total night charge      3333 non-null   float64
 15  total intl minutes      3333 non-null   float64
 16  total intl calls        3333 non-null   int64
 17  total intl charge       3333 non-null   float64
 18  customer service calls  3333 non-null   int64
 19  churn                   3333 non-null   bool
dtypes: bool(1), float64(8), int64(8), object(3)
memory usage: 524.0+ KB
```

Exploratory Data Analysis

- SMOTE utilization balancing out minority value in target
- 3 categorical features:
  - Churn, VM plan, INT'L plan
- Feature Engineering:
  - Dummy variables,
  - ROC AUC accuracy pre-Gridsearch CV

```
Counter({0.0: 2850, 1.0: 483})
Counter({0.0: 2850, 1.0: 2850})
```



Models

Final Model- Decision Trees--Base(Log Regression)

Accuracy: 91% | 75%

Model effectively learned from synthetic data during SMOTE

Chosen based on high accuracy predictions.

Final model is overfit and cannot accurately predict new information on an average.

Summary

Bagged Trees- ⅔ of data is split and left aside while other ⅓ is OOB and used as test set. Model creates *X* number of decision trees trained on *X* number of bootstrapped training sets. Final value is average of all *X* decision trees.

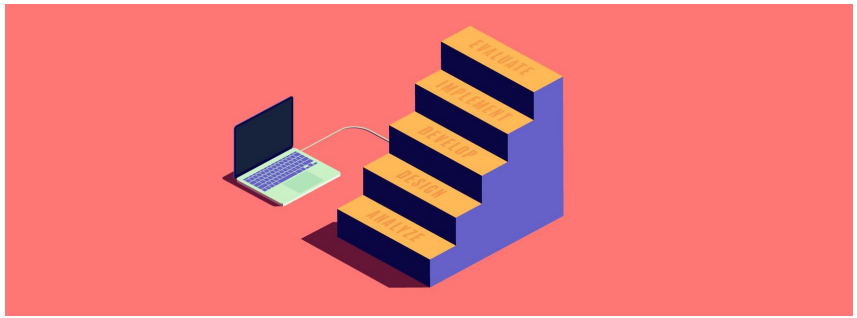
Random Forest- improves on bagging by decorrelating trees with intro of splitting random subset features. This causes variance to be averaged away.

XG Boost- high performance implementation of gradient boosted decision trees.
Language: C++, Easy to implement.

Logistic Regression- doesn't need linear relationship between target and dependant variables. Dependant variables MUST be independent of each other. MUST have little to no multicollinearity. Independent variable must be linearly related to log odds.

Next Steps??

- Further research suggests there are other models that can also be utilized (Neural Networks, Ridge Regression)
- Continuous variables could be separated into categorical for more effective decision tree modeling.



Thank you!

GitHub Repository

Linkedin Profile

Email: zacharyhyde14@gmail.com

Resources:
SMOTE, Random Forest, XGBoost, SciKit-Learn