Telco Churn Prediction with Big Data Paper Summary

Tamer Tahamoqa 20208019

(Authors used Apache Hadoop and Apache Spark)

Abstract

* Authors narrowed down their big data perspective by only utilizing the 3 standard V’s of big data (Volume, Variety, Velocity).
* Prediction performance has been significantly improved by using a large volume of training data, a large variety of features from both Business Support Systems (BSS) and Operations Support Systems (OSS), and a high velocity of processing new incoming data.
* The authors deployed the churn prediction system in one of the biggest mobile operators in China.
* The system can provide a list of prepaid customers who would be most likely to churn in the upcoming month from millions of active customers, with 96% precision for the top 50,000 predicted churners in the list. The previous deployed system had 68% precision.
* The main **value** out of this system would be automatic matching of retention campaigns with the predicted potential churners to significantly boost their recharge rates.

Introduction

* The used dataset is a 9-month dataset from around two million prepaid customers in one of the biggest operators in China.
* Operations Support Systems (OSS) data composed 97% of the size of the telco operator data assets.
* The authors also integrated churn prediction with the retention campaign systems as a closed loop; after each campaign the churners who would have accepted the retention offers or declined them would be known and as such can be labeled for training a multi-class classifier to match proper retention offers to each potential churner.
* The authors were trying to provide an answer to the following question: “How to monetize telco big data?”

Previous Work

* In the telco industry, data-driven churn predictive modeling generally includes constructing useful features and building good classifiers.
* Most customer behavior features are extracted from Business Support Systems (BSS).
* A wide variety of classifiers were adopted: Logistic Regression, Decision Tree, Random Forest, Boosting Algorithms (variants of adaboost), Boosted Trees (gradient-boosted decision trees), Neural Network, Evolutionary Computation algorithms (e.g: Genetic Algorithm, Ant Colony Optimization), Ensemble of Support Vector Machines, Ensemble of Hybrid methods.
* The authors claim previous works have two limiting factors:
  + Sophisticated techniques have been applied in churn prediction for years and it is getting harder to make a breakthrough; which motivates seeking opportunities using telco big data. Also, although Operations Support Systems (OSS) data makes 97% of most telco data assets; they were rarely studied before.
  + Matching retention campaigns with potential churners has not been investigated to maximize the overall profit, which causes limited business value of existing churn prediction models.

Telco Big Data Platform

* Operations Support Systems (OSS) generally support network management, network inventory, service provisioning, network configuration, and fault management.
* Business Support Systems (BSS) generally support the following processes: product management, order management, revenue management, customer management.
* OSS and BSS systems often work separately.
* Telco data storage requires big data computing platforms.
* Data volume for OSS data was around 2.2 Terabytes per day for the operator the authors worked with.
* BSS data consisted of data from around 140 tables.
* The authors used the Hadoop File System (HDFS) with Apache Hive and Apache Spark for feature engineering purposes.
* Classification components were coded in Apache Spark.

Prediction and Retention

* The authors heavily used Spark SQL for feature engineering instead of Apache Hive since Spark SQL is much faster than Apache Hive.
* Intermediate results during feature engineering were stored as Hive tables.
* The authors chose the Random Forest classifier for the system since it yielded the best predictive performance among several widely-used classifiers (Randorm Forest, LIBFM, GBRT, LIBLINEAR (l2-regularized logistic regression)).
* The authors also chose the Random Forest classifier for the retention systems.

Results and Conclusions

* The training dataset was imbalanced, the authors used the following methods to deal with this issue during training: Unbalanced training, Up Sampling, Down Sampling, Weighted Instance (Weighted Loss). **Weighted instance** gave the best results.
* Random Forest yielded the best results and was chosen for both churn prediction and retention prediction.
* The authors claim that more utilization of OSS data with BSS data with more volume, variety, and velocity (utilization of latest data) would improve performance and would offer more value to telco companies.