Group 1 Project Report
Customer Churn Prediction Model using machine learning in big data platform

Submitted by:
Rashad Dana
Mohammad Abu Halib
Dana Estetieh
Tamer Tahamoqa

Motivation

One of the challenges that any telecom company faces is customers' churn, due to its direct effect on the revenues. Building a predictive churn model helps the organization make proactive changes to their retention efforts that drive down churn rates. By using machine learning algorithms it would be possible to analyze and predict churner and non-churner customers. We have focused our research and implementation on the Spark framework using Python and the PySpark Wrapper. The implemented classifier models came from the ML library in PySpark.

Dataset

A telecommunication churn data set obtained from kaggle.com was used with the name *'telco-customer-churn'* [1]. The data set contains 7043 rows and 21 variables including the churn outcome. The data was split in different ratios for the training and testing purposes for different model training experiments.

Contributions

TAMER TAHAMOQA

Readings

- Yiqing Huang, Fangzhou Zhu, Mingxuan Yuan, Ke Deng, Yanhua Li, Bing Ni, Wenyuan Dai, Qiang Yang, and Jia Zeng. 2015. Telco Churn Prediction with Big Data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (*SIGMOD '15*). Association for Computing Machinery, New York, NY, USA, 607–618. DOI:https://doi.org/10.1145/2723372.2742794 [2]

- Wassouf, W.N., Alkhatib, R., Salloum, K. *et al.* Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. *J Big Data* **7,** 29 (2020). https://doi.org/10.1186/s40537-020-00290-0 [3]

- Nurul Izzati Mohammad, Saiful Adli Ismail, Mohd Nazri Kama, Othman Mohd Yusop, and Azri Azmi. 2019. Customer Churn Prediction In Telecommunication Industry Using Machine Learning Classifiers. In *Proceedings of the 3rd International Conference on*

Vision, Image and Signal Processing (*ICVISP 2019*). Association for Computing Machinery, New York, NY, USA, Article 34, 1–7. DOI:https://doi.org/10.1145/3387168.3387219 [4]

Work

- Model training and tuning with Exhaustive Hyperparameter Grid Search and Cross Validation.
- ML Models: Decision Tree Classifier, Random Forest Classifier, GBT Classifier.

RASHAD DANA

Readings

- Olayemi Olasehinde. 2018. Computational Efficiency Analysis of Customer Attrition Prediction Using Spark and Caret Random Forest Classifier. Journal of Information & Knowledge Management · September 2018 [5]
- Muhammad Joolfoo, Khalid Joolfoo. 2020. Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform. Journal of Critical Reviews 7(11):1991 [6]
- Hend Sayed et al. 2018. Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib Packages: A Comparative Study. International Journal of Advanced Computer Science and Applications 9(11) [7]

Work

- Data preprocessing: Features scaling, PCA dimensionality reduction.
- ML Models Implementation: Random Forest Classifier, Logistic Regression, GBT Classifier and Linear SVC.

MOHAMMED ABUHALIB

Reading

- X. Hu, Y. Yang, L. Chen and S. Zhu, "Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network," 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2020, pp. 129-132, doi: 10.1109/ICCCBDA49378.2020.9095611.
- L. Butgereit, "Work Towards Using Micro-services to Build a Data Pipeline for Machine Learning Applications: A Case Study in Predicting Customer Churn," 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE), Aswan, Egypt, 2020, pp. 87-91, doi: 10.1109/ITCE48509.2020.9047807.

Work

- Data preprocessing: indexers, one hot encoder, and Vector Assembler using Pipeline.
- ML Models: Decision Tree.

DANA ESTETIEH

Reading

- Abdelrahim Kasem Ahmad , Assef Jafar and Kadan Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform” Faculty of Information Technology, Higher Institute for Applied Sciences and Technology, Damascus, Syria

- Sachin Bhoite, “Customer Churn Analysis and Prediction”- International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 363-366, 2019, ISSN:-2319–8656

Work

The first paper has been summarized, see appendix.

Literature Review

TAMER TAHAMOQA

Yiqing Hunag et al. [2] deployed a churn prediction system in one of the biggest mobile operators in China. The prediction performance of the system significantly outperformed the previous deployed system, the previous deployed system had 68% precision for the top 50,000 predicted churners while the new system achieved 96% precision. The system achieved the better results 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 used a 9-month dataset from around two million prepaid customers, OSS data composed 97% of the size of the mobile operator’s data assets. The authors noted that although OSS data makes 97% of most telco operator data assets; they were rarely studied before. The authors also integrated churn prediction with 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 customer.

The authors used Apache Hadoop for storage, Apache Hive and Apache Spark for feature engineering. Training the classifier models was conducted in Apache Spark. Data volume for the OSS data was around 2.2 terabytes per day, and the BSS data consisted of data from around 140 tables. Since the training data was imbalance; the authors experimented with unbalanced training, up sampling, down sampling, and weighted instance (weighted loss). Weighted instance gave the best results. The best classifier that yielded the best results was the Random Forest classifier 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 offer more value to telco companies.

Wissam Wassouf et al. [3] implemented a customer loyalty prediction system at the Syriatel Telecom Company, the authors used a dataset supplied by Syriatel that contained 127 million records and 220 features for training and testing. Customers were segmented based on the Time-Frequency-Monetary (TFM) approach which is adjusted from the Recency-Frequency-

Monetary (RFM) approach, the level of loyalty was then defined for each segment or group. The loyalty level descriptors were taken as categories, choosing the best behavioral features for customers, their demographic information such as age, gender, and the services they share. Several classification algorithms were applied based on the descriptors and chosen features to classify new users by loyalty.
The authors used the Hortonworks Data Platform (HDP) Hadoop framework, Apache Hadoop was used for data storage and Apache Spark was used for most phases of data processing and classifier model training on Spark ML, Apache Yarn for resource management, Apache Zeppelin as a development user interface, Apache Ambari for system monitoring, Apache Ranger for system security, Apache Flume and Apache Sqoop for data acquisition from Syriatel data sources to Hadoop, and Apache Hive for processing structured and semi-structured data. The Gradient-boosted tree (GBT) classifier was found to be the best in the binary classification task of (loyal vs unloyal) and the Random Forest classifier was found to be the best in the multi-classification task of (very low loyalty, low loyalty, medium loyalty, high loyalty, very high loyalty).

Nurul Izzati Mohammad et al. [4] utilized the *'telco-customer-churn'* dataset from kaggle [1] which is also used in this project to train customer churn classifiers. The authors split the data into a 70/30 train/test split and applied dummy variables to the categorical attributes. The authors conducted exploratory analysis of the data and found that 73.4% of the dataset was composed of customers who stayed loyal with the service provider and 26.6% of customers who churned. The range of tenure duration for the dataset was between two months and 72 months. The longer contracts had the lowest churn rates; two-year contracts had 3% churns, the one-year contracts had 11% churns, and the monthly contracts had 43% churns.

The authors used Pandas dataframes and the scikit-learn machine learning package for their implementation, the authors conducted classifier model training experiments with Logistic Regression, Multilayer Perceptron Artificial Neural Network (ANN), and the Random Forest classifier. The classifier models were evaluated by accuracy, precision, recall, and error rate to find the best classifier. The authors found that the most important features that influenced the prediction of the models were the total charges, monthly contract, and fiber optic internet service features. The authors have found that applying recursive feature elimination (RFE) for feature selection significantly improved results; the Logistic Regression classifier managed to achieve 100% accuracy, precision, and recall on the test set.

RASHAD DANA

Olasehinde (2018) used spark to analyze the performance of the Random Forest Models from both SparkML and the CARET package from the R language. While the accuracies scored were observed to be very close 0.7982018 for the CARET model and 0.7934272 for the model from the Apache Spark package, the precision of each model varied significantly with the model from spark package and CARET scoring 0.7825538 and 0.9115646 respectively. The paper also showed the degree of which each variable has influenced the prediction of the models.

Muhammad Joolfoo et al (2020) researched customer churn prediction in telecom industry using machine learning and big data platform (Hortonworks). With Spark being their execution engine processing and KNN algorithm as the main focus for implementing the prediction task. Logistic Regression was also implemented for comparison.
Accuracy, recall, precision and f-score where the 4 measures used to evaluate the performance of each algorithm. It was concluded that KNN has accuracy of 80% and area under the curve of 71%.

Hend Sayed et al (2018) conducted a performance comparison between the ML and MLlib libraries in Spark on a banking customers churn dataset and a decision tree algorithm. The conclusion reached from their study that the newer Library (ML) performed better regarding the accuracy of predictions than the MLlib 0.79 and 0.73 respectively. While the ML library took more time (25 seconds) to apply data transformations and train the model than the MLlib (6 seconds), it needed only 5 seconds to evaluate the testing data versus 14 seconds for MLlib.

MOHAMMED ABUHALIB

The authors design a combined prediction model based on two models of decision tree and neural network, to predict customer churn in a supermarket. By comparing the prediction accuracy of the three models, the validity of the combined prediction model is verified, The decision tree prediction process is performed in two steps; Build and evolve a decision tree using the training set (⅔) (⅓) and test the attribute values of each node, classify the input data, and use the attribute values of this class to complete the estimation of the prediction object. the decision tree model uses the bootstrap method to improve the accuracy of the algorithm. For Neural Network Customer Churn Prediction Model; The model structure of a typical neural network includes an input layer, hidden layer, and output layer, which are connected by several neurons. The biggest advantage of the combined customer churn prediction model is that it can integrate the results of the two models to clearly distinguish between churn customers and non-churn customers, and for customers in between. The empirical results show that the combined prediction model can not only have a better interpretation ability like a decision tree model, but also a higher prediction accuracy rate of a neural network model, which can better make up for the shortcomings of a single prediction model and can also get more stable and accurate prediction results

The paper was done at a public company listed on the Johannesburg Stock Exchange in South Africa. The author mainly focuses on the management of the data pipeline which feeds the machine learning algorithms and the management of the results of the neural network. The author also describes work on creating and maintaining the data flow using microservices. The author created the Data Pipeline used Spring with Netflix Eureka registry. Each major section was developed as a REST application. Each section started the next REST application and did not wait for a return value. The data pipeline contains the following sections: data sources, Re-format data, Train/Test live data split, Train /Test MLP, process recent data and Result.

DANA ESTETIEH

Ahmad et al. “proposed a model for customer churn prediction in the big data platform. The author used Social Network Analysis (SNA) and Area Under Curve (AUC). Four tree based algorithms were chosen Decision Tree, Random Forest, Gradient Boosted Machine Tree “GBM” and Extreme Gradient Boosting “XGBOOST”. The model was trained, tested and evaluated in addition

to applied feature engineering. XGBOOST tree model achieved the best result with AUC value 93.301%”

Sachin Bhoite”a prediction model was built based on old data which was about customers that already churned by using different algorithms, and detected which algorithm was more accurate. The result was Logistic regression algorithm gave higher accuracy 80.38%, Decision Tree 77.81% , Random Forest Tree , 80.02%”

Proof of Concept implementation

MOHAMMED ABUHALIB

Data Preprocessing:

In data Preprocessing, the dataset transformed and normalized to feeding it to the Machine learning model. Vector Assembler and one-hot encoding were used to vectorize the features and the indexers for label indexing. Spark Pipeline used to fit the dataset one time.

Modeling:

The data set was randomly split in an 80/20 train/test split with a set random seed value and feed to the Decision Tree classification. 0.79 accuracy and 0.20 test error were obtained from decision tree classification.

RASHAD DANA

Data Preprocessing:

Based on Mohammad’s work, the data has been scaled according to the MinMax method, then PCA dimensionality reduction has been applied in the following ways:

- Dimensionality reduction on unscaled data (pcaFeatures)
- dimensionality reduction on scaled data (pcaScaledFeatures)
- Scaling of dimensionally reduced unscaled data (scaledPcaFeatures)

The purpose was to test the different combinations of dimensionality reduction and data scaling on various ML models.

Modeling:

Implemented Random Forest Classifier, Logistic Regression, GBT Classifier and Linear SVC. The data has been split in 0.7, 0.3 ratios for training and testing respectively for both the GBT

Classifier and the Linear SVC. On the other hand, 0.8, 0.3 was taken for the Random Forest Classifier and Logistic Regression.
A preliminary run, before parameters tuning, showed that the algorithms with the higher accuracies are the GBTClassifier (0.801728) and the Random Forest Classifier (0.800143).

TAMER TAHAMOQA

Two training experiments were conducted with the following classifier models in PySpark ML: Decision Tree classifier, Random Forest Classifier, Gradient-Boosted Tree (GBT) classifier. The data set was randomly split in a 70/30 train/test split with a set random seed value for reproducibility. Exhaustive Hyperparameter Grid Search model training with 10-fold Cross Validation on the training set were conducted for each model on every experiment whereby all available cpu cores would be used for parallel training and the model with the best average accuracy metrics on the validation folds would be retrieved for testing on the test set. It is worth noting that the *spark.driver.memory* of the Spark Session had to be increased to handle Java Heap Out of Memory exceptions due to the parallel training consuming significant RAM memory.

The following metrics were used to measure the performance of the classifier models on the test set: Accuracy, F1-Score, Weighted Precision, Weighted Recall, Repeater Operating Characteristic Area Under Curve (ROC AUC).

The first experiment had no PCA dimensionality reduction and the preprocessing steps implemented by Rashad Dana and Mohammad Abu Halib were used, the best performing Random Forest classifier had the highest accuracy and Weighted Recall scores while the best performing Gradient-Boosted Tree classifier had the highest F1-Score, Weighted Precision, and ROC AUC scores (**see Table 1**).

The second experiment had PCA dimensionality reduction implemented by Rashad Dana with the number of principal components (k) being set to seven components. The number was arbitrarily chosen as to our knowledge there doesn't exist a way to set the number of optimal components based on the percentage of explained variance in PySpark. All the new PCA columns by Rashad Dana were used for experiment 2 (**see Table 2**).

Classifier Model	**Accuracy**	**F1-Score**	**Weighted Precision**	**Weighted Recall**	**ROC Area Under Curve**
Decision Tree	77.53%	77.12%	76.83%	77.53%	0.6914
Random Forest	**78.53%**	77.14%	76.98%	**78.53%**	0.6716

Gradient-Boosted Tree (GBT)	78.35%	**77.74%**	**77.41%**	78.35%	**0.6941**

Table 1: Experiment 1 results (no PCA dimensionality reduction)

Classifier Model	Accuracy	F1-Score	Weighted Precision	Weighted Recall	ROC Area Under Curve
Decision Tree	78.03%	77.91%	77.81%	78.03%	0.7095
Random Forest	80.16%	79.18%	79.02%	80.16%	0.7024
Gradient-Boosted Tree (GBT)	**80.20%**	**79.26%**	**79.08%**	**80.21%**	**0.7039**

Table 2: Experiment 2 results (with PCA dimensionality reduction)

Unfortunately, to the best of our knowledge, we have found that the recursive feature elimination (RFE) feature selection method that was used by Nurul Izzati Mohammad et al. [4] on the kaggle dataset to achieve 100% accuracy using the Logistic Regression classifier to not be available on Apache Spark. As such, we have not conducted an experiment utilizing this feature selection method.

Limitations

- Small number of data samples.
- Lack of domain knowledge required for effective feature engineering.
- Lack of time required to conduct more classifier model training experiments.

Future Work

- Training a Multilayer Perceptron (MLP) classifier model.
- Testing using an ensemble of good classifier models while trying multiple prediction averaging techniques.

References

[1] 'telco-customer-churn' Kaggle Dataset: https://www.kaggle.com/blastchar/telco-customer-churn

[2] Yiqing Huang, Fangzhou Zhu, Mingxuan Yuan, Ke Deng, Yanhua Li, Bing Ni, Wenyuan Dai, Qiang Yang, and Jia Zeng. 2015. Telco Churn Prediction with Big Data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (*SIGMOD '15*). Association for Computing Machinery, New York, NY, USA, 607–618. DOI: https://doi.org/10.1145/2723372.2742794

[3] Wassouf, W.N., Alkhatib, R., Salloum, K. *et al.* Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. *J Big Data* **7,** 29 (2020). https://doi.org/10.1186/s40537-020-00290-0

[4] Nurul Izzati Mohammad, Saiful Adli Ismail, Mohd Nazri Kama, Othman Mohd Yusop, and Azri Azmi. 2019. Customer Churn Prediction In Telecommunication Industry Using Machine Learning Classifiers. In *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing* (*ICVISP 2019*). Association for Computing Machinery, New York, NY, USA, Article 34, 1–7. DOI: https://doi.org/10.1145/3387168.3387219

[5] Olasehinde, Olayemi & Victor, Olanrewaju & Fakoya, Johnson. (2018). Computational Efficiency Analysis of Customer Attrition Prediction Using Spark and Caret Random Forest Classifier. Journal of Information & Knowledge Management. 8. 8-16.

[6] Joolfoo, Muhammad & Jugurnauth, Rameshwar & Joolfoo, Khalid. (2020). Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform. Journal of Critical Reviews. 7. 1991. 10.31838/jcr.07.11.308.

[7] Sayed, Hend & Abdel-Fattah, Manal & Kholeif, Sherif. (2018). Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib Packages: A Comparative Study. International Journal of Advanced Computer Science and Applications. 9. 10.14569/IJACSA.2018.091196.

Appendix

Introduction:

Customer churn is one of the most important metrics for a growing business to evaluate, due to its effect on the revenues of the companies. Customer churn can prove to be a roadblock for an exponentially growing organization and a retention strategy should be decided in order to avoid an increase in customer churn rates.

The churn rate is a particularly useful measurement in the telecommunications industry. As most customers have multiple options from which to choose, the churn rate helps a company determine how it is measuring up to its competitors.

By using machine learning we can analyze, predict the way customer respond to these services [1].

The aim is to develop a customer churn prediction model by using many algorithms through Spark environment by working on a large dataset and comparing the results of four trees based machine learning algorithms.

Dataset

The size of the data was more than 70 terabytes, it contains transactions for all customers during nine months before the prediction baseline and it's related to calls, SMS, the internet, complaints, and others.The volume is about 70 Terabyte and it has different data formats which are structured, semi-structured, and unstructured. The data has been classified as per its types.

Some challenges faced during collecting and analyzing the data such as missing data, unbalanced data, and others.

Many frameworks were installed to go through all phases such as Hortonworks Data Platform (HDP)[11], Hadoop Distributed File System to store the data (HDFS)[10], Spark engine to process the data [9], Yarn to manage the resources and others [8]. Some of them are Apache Flume was used to transfer unstructured and semi-structured data from outside SYTL-BD into HDFS and Apache SQOOP was used to transfer the bulk of data between HDFS and relational databases (Structured data) by using Map jobs [2,3,4].

In order to get the best result with feature engineering and data exploration tasks, parquet file type was the chosen format type [3].

The data was processed to convert it from its raw status into features to be used in machine learning algorithms by using Spark engine for both statistical and social features. The library used for SNA features is the Graph Frame.

As mentioned above, one of the challenges is missing values. The preferable method chosen to deal with is filling out the missing values with other values derived from either the same features or other features.

Below are the changes that applied to the data:

- Records that contain more than 90% of missing features were deleted.
- Features that have more than 70% of missing values were deleted.
- For the missing categories in categorical features, they were replaced by a new category called ‘Other’.
- The missing numerical values were replaced with the average of the feature.
- The number of categorical features were 78, the first 31 most frequent categories were chosen and the remaining categories were replaced with a new category, so the total number is 32 categories.
- There are some other features with a numeric character but they contain only a limited number of duplicate values in more than one record. This indicates that they are categorical so we have dealt with them as categorical features, but the experiment shows that they perform worse with the model, so that they have been deleted.

The data was divided into two groups: training group 70% and testing group 30%. The training sample size is 420,000 [6,7]. Four algorithms were trained and followed the same steps, best value after multiple experiments for each algorithm is as follows:

- Decision tree: 398 nodes in the tree and the depth value was 20.
- Random Forest: 200 trees.
- GBM: 200 trees (gave the best result than RF and DT).
- XGBOOST: 180 trees.

Results and discussion

The results were analyzed to compare the performance regarding the different sizes of training data. It increased significantly by adding SNA features with the statistical features to the classification algorithms, where the max reached value of area under the curve (AUC) was 93.3%. Below table (2) shows comparing AUC results before and after adding SNA to statistical features:

To deal with unbalanced data, it was found that XGBOOST and GBM algorithms gave the best performance without any rebalancing techniques, while Random Forest and Decision Tree algorithms gave a higher performance by using undersampling techniques.

After applying the four algorithms, XGBOOST has been chosen to be the classification algorithm in this proposed predictive with an AUC value of 93.3%, the second place is GBM algorithm with an AUC value of 90.89%, the third place is Random Forest with an AUC value of 87.76%, and Decision Trees came last in AUC ranking with values of 83%.

By adding the Social Network Analysis features changed the ranking of the important features as follows:

1. The first feature is the time period between moving from the current community to the other operator's GSM.

2. The second feature is Days of Last Outgoing transaction.

3. The third important feature is total balance since most churners had low balance compared with the active customers.

4. The fourth feature is Average of Radio Access Type where most of the churners had more 2G internet sessions than 3G sessions.

5. The fifth feature in importance is Local Cluster Coefficient, where the customers with very low LCC value are less likely to churn.

6. The sixth feature is the Percentage of Transactions to/from other Operator as it represents the effect of friends on the churn decision.

7. The last feature in importance is Customer's Age, where they found the customers who are less than 32 years old have more likelihood to churn than the others.

The method of preparation and selection of features and entering the mobile social network features had the biggest impact on the success of this model, since the value of AUC in SyriaTel reached 93.301%.

SNA increased the performance of the churn prediction model, since they gave a different insight to the customer from the social point of view.

The system has been evaluated and tested on all prepaid SyriaTel customers. The dataset was divided into (Offered, NotOffered), Offered means proactive action to retain the customers who are predicted to leave and NotOffered dataset left without any action. The results were very good and the best AUC value was 89% for XGBOOST on "NotOffered", and 47% of customers were retained from the Offered dataset. This confirms the accuracy of the models and the great impact on the organization as the revenue will be increased and the churn rate will be decreased by about 1.5%.

References:

1. International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 363-366, 2019, ISSN:-2319–8656

2. https ://spark .apach e.org/docs/lates t/sql-progr ammin g-guide .html.

3. https ://parqu et.apach e.org/.

4. https ://avro.apach e.org/.

5. Name of this article Ahmad et al. J Big Data (2019) 6:28 https://doi.org/10.1186/s40537-019-0191-6

6. Burez D, den Poel V. Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009;36(3):4626–36.

7. Xie J, Rojkova V, Pal S, Coggeshall S. A combination of boosting and bagging for kdd cup 2009—fast scoring on a large database. J Mach Learn Res Proc Track. 2009;7:35–43.

8. https ://hadoo p.apach e.org/docs/curre nt/hadoo p-yarn/hadoo p-yarn-site/YARN.html.

9. https ://spark .apach e.org/.

10. https ://hadoo p.apach e.org/docs/r1.2.1/hdfs_desig n.html.

11. https ://horto nwork s.com/.