Predictive Analytics Using Big Data for Increased Customer Loyalty Syriatel Telecom Company Case Study Paper Summary

Tamer Tahamoqa 20208019

(Authors used Hortonworks Data Platform: Hadoop ecosystem with Apache Spark)

Abstract (with notes in the Related work section)

* Emergence of big data concepts introduced a new wave of Customer Relationship Management (CRM) strategies.
* Big Data analysis would help to describe customer behavior, understand customer habits, develop appropriate marketing plans for organizations to identify sale transactions, and build long-term loyalty relationships.
* The authors used a dataset that contained 127 million records for training and testing supplied by Syriatel corporation and contained 220 features.
* Customers were segmented based on the Time-frequency-monetary (**TFM**) approach which is adjusted from the Recency-frequency-monetary (RFM) approach, where Time (T): total of call duration and Internet sessions in a certain period of time, Frequency (F): services used frequently within a certain period, Monetary (M): The money spent during a certain period. 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 Gradient-boosted tree was found to be the best in binary classification, and the Random Forest classifier was found to be the best in multi-classification.

Introduction

* The authors claim that the cost associated with customer gain is usually higher than the cost associated with maintaining a customer; therefore the telecom sector would greatly benefit from predicting the amount of income they may receive from their active customers.

Research objectives

* Customer value was analyzed by TFM segmentation and the level of loyalty per each segment was determined.
* A set of features was derived from the telecom data.
* The following classification models were trained on the chosen features and level of loyalty per each segment: Random Forest, Decision Tree, Gradient-boosted Tree classifier, Multilayer perceptron (MLPC) with 4 layers and 5 nodes per layer.
* The most accurate model was selected based on several criteria.
* The model was used to classify new users by loyalty, and to derive the rules that showed the characteristics of each level of loyalty.

Research tools (HDP)

* The authors used the Hortonworks Data Platform open-source framework (HDP), the authors used the Hadoop HDFS for data storage and Apache Spark for most phases of data processing and classifier model training, Yarn for resource management, Zeppelin as a development user interface, Ambari for system monitoring, Ranger for system security, Flume and Sqoop for data acquisition from Syriatel data sources to HDFS, Hive for processing structured and semi-structured data.
* The hardware resources were composed of 12 nodes with 32 GB of RAM, 10 TB storage-capacity, and 16-core processors per node.

Data preparation

* The authors collected the Syriatel data to the Hadoop environment.
* The authors used the Scala language to perform data preparation, attribute extraction, model training and testing.
* The authors used the **Spark ML** library to train the classifier models.
* The data was divided into a 70/30 train/test split.
* Attributes who at least contained 70% missing values were deleted.
* Missing numerical values were imputed with mean values for each attribute.

Results and discussion

* For binary classification (Loyal, not Loyal): the Gradient-boosted Tree model yielded the best results: Accuray 87%, Precision: 76%, Recall 65%, f1-score: 70%, with lower scores associated with classifying the ‘Loyal’ class.
* For multi-class classification (Very low loyalty, low loyalty, medium loyalty, high loyalty, very high loyalty): The Random Forest classifier yielded the best results: Accuracy: 74%, Weighted Average Precision: 73%, Weighted Average Recall: 74%, Weighted Average F1-score: 72%
  + The ‘Low loyalty’ class had the worst performance metric values.
  + Note: the authors did not report the macro average performance metrics, I think because of the low performance on the ‘low loyalty’ class the metrics were low so they reported the weighted average metrics.
  + Note: the authors could not use the Gradient-boosted Tree model for multi-class classification as it only supported binary classification.