Customer Churn Prediction In Telecommunication

Industry Using Machine Learning Classifiers Paper Summary

Tamer Tahamoqa 20208019

(The authors used Pandas dataframes and scikit-learn instead of Apache Spark they have used the Kaggle Telco Customer Churn [dataset](https://www.kaggle.com/blastchar/telco-customer-churn/notebooks))

Abstract

* The authors aimed to identify the factors that influence customer churn and develop an effective churn prediction model.
* The authors used the publicly available Kaggle Telco Customer Churn dataset with 7043 records with 21 attributes.
* The authors chose the following classifier algorithms: Logistic Regression, Artificial Neural Network (ANN) or Multilayer Perceptron, Random Forest.
* The classifier models were evaluated by accuracy, precision, recal, and error rate to find the best classifier.
* **Logistic Regression** outperformed the other classifiers.

Data Preparation

* The authors split the data into an 70/30 train/test split.
* The authors applied dummy variables to categorical attributes.

Experiment Results and Analysis

* 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% churn, the monthly contracts had 43% churns.
* The most important features that influenced the prediction of the models were total charges, monthly contract, and fiber optic internet service.
* **Applying feature selection recursive feature elimination (RFE) significantly improved results**, the method removes the weakest features and reduces overfitting.
  + **Logistic Regression**:
    - Without RFE:
      * Accuracy: 80.05%
      * ROC AUC: 0.73
    - With RFE:
      * Accuracy: 100%
      * ROC AUC: 1.0
      * Precision: 100%
      * Recall: 100%
  + Artificial Neural Network:
    - Without RFE:
      * Accuracy: 80.28%
      * ROC AUC: 0.72
    - With RFE:
      * Accuracy: 85.55%
      * ROC AUC: 0.77
      * Precision (Non-churn): 80%
      * Recall (Non-churn): 92%
      * Precision (Churn): 70%
      * Recall (Churn): 62%
  + Random Forest:
    - Without RFE:
      * Accuracy: 77.73%
      * ROC AUC: 0.69
    - With RFE:
      * Accuracy: 98.44%
      * ROC AUC: 0.96
      * Precision (non-churn): 100%
      * Recall (non-churn): 99%
      * Precision (Churn): 100%
      * Recall (Churn): 93%