Predicting Customer Churn Using Big Data Analytics

Abstract

Customer churn is a significant concern for businesses, affecting growth and profitability. This project seeks to address this issue through the development of a complete churn prediction system within a big data analytics framework. The three big data technologies used in this project are Hadoop for distributed text (raw) storage and the initial data preparation step, PySpark for distributed data analysis/scalable machine learning, and ZooKeeper for managing complex distributed systems. The end-goal of this project is to design and deploy a predictive model that locates customers who are at high-risk of churning. Businesses can then leverage advanced models like this to intervene and take action to retain customers who would have typically left and switch to a competitor. This report will outline the approach, technology used, and collaboration of the team members.

Introduction

Customer churn, or the percentage of customers who have stopped doing business with a service/product, is critically important to companies who wish to maintain a customer base and continue to grow. Analyzing customer churn is commonly done through various traditional methods. These methods often lend themselves towards limitations of sample size or complexity of variables. The aim of this project is to use today's technologically advanced big data to enhance customer churn ecological validity, rather than using one or two variables to predict customer churn.

Goals

- Develop and validate predictive models for customer churn.
- Utilize Hadoop to store large amounts of structured and unstructured data and, once stored in Hadoop's HDFS, begin the process of analyzing data.
- Utilize PySpark to analyze customer lists and build workflows for actionable analytics.
- Use ZooKeeper to administer and supervise the real-time, dynamic executions of all data preparation and exploratory data analysis.

Team Members

- Shashank A Bhat (220962008)
- Abhijith Pai (220962006)
- Tejas Patil (220962002)
- Umair Ismail (220962266)

Contribution of Each Team Member

Tejas M Patil:

- Data Management and EDA: Led the collection and integration of data into Hadoop’s HDFS and performed initial exploratory data analysis to identify key patterns for feature engineering.
- Model Evaluation: Contributed to evaluating machine learning models by assessing their performance and conducting cross-validation to ensure stability and robustness.

Shashank A Bhat:

- Feature Engineering: Responsible for extracting and selecting features that are relevant to churn prediction, which involved transforming raw data into usable features.
- Model Development: Achieved development and tuning of machine learning models using PySpark, including logistic regression, decision trees, and ensemble approaches..

Abhijith Pai:

- Data cleaning/preprocessing: Analyzed the process of data cleaning to impute null values and validate data integrity, which improved data quality for analysis.
- ZooKeeper Integration: Involved with ZooKeeper integration to manage system coordination with real-time processing of data.

Umair Ismail:

- System Architecture Design**:** Contributed to designing the overall system architecture, ensuring effective integration of Hadoop, PySpark, and ZooKeeper.
- Testing and Debugging**:** Participated in testing the data processing pipelines and machine learning models, identifying and resolving issues to ensure smooth operation.