Raksa Ma

**Final Report: Orange Telecom Customer Churn Prediction**

1. **Introduction**

In telecommunications industry, customer burn poses a significant risk to business sustainability and growth. Churn is defined as the phenomenon where customers discontinue their subscription or service, often moving competitors. Predicting customer churn is important for telecom companies, as retaining existing customers is more cost-effective than acquire new customers (Hadden at al., 2007). Companies can implement target retention strategies by fining customers at risk of churning, therefore reducing revenue loss and improve customer satisfaction.

Big data and machine learning has opened new opportunities for solving business challenges including churn prediction. Decision trees, support vector machines, and neural networks have demonstrated success in predicting churn across various industries (Verbeke et al., 2012; Burez & Van den Poel, 2009). This project explores PySpark application as a distributed data processing framework, to build a predictive model for customer churn using Orange Telecom dataset on Kaggle. The datasets include customer demographics, service usage patterns, customer interaction data, perfect for creating robust predictive models.

The first goal of this project is to develop a machine leaning pipeline for churn prediction. Secondly, the goal is to find insights into factors contributing to customer churn. Findings from this experiment aim to provide actional insights that can be used to improve customer retention strategy for Orange Telecom. Finally, this case study applies PySpark capabilities to solve real-world problems, reinforcing its importance of distributed computing for large-scale data analysis.

1. **Background and Related Work**

Customer churn prediction has been considered as an important research area in customer relationship management (CRM) in industries such as telecommunications, banking and retail in which maintain customer loyalty is crucial. Features such as customer behaviors, service usage pattern, and demographic information are used by predictive model to identify customer at risk of canceling their subscription. These insights allow businesses to take proactive measures such as offering personalized incentives and improve service quality to reduce churn rate (Ngai et al., 2009).

* 1. **Machine Learning in Churn Prediction**

Machine learning has been used in churn prediction by using techniques that could handle large-scale, high-dimensional datasets and uncover complex relationships between features. Traditional methods such as logistic regression have been used for simplicity and interpretability, but they often fall short in capturing non-linear relationships (Witten et al., 2016). More advanced approaches such as decision trees, support vector machines (SVMs), random forests, and gradient boosting have demonstrated improved predictive performance (Burez & Van den Poel, 2009; Verbeke et al., 2012).

Deep learning techniques have gained popularity in recent years due to their capability to extract features and model non-linear relationships automatically. However, their computational demands and lack of interpretability often pose challenges, particularly for business seeking actional insights (Chawla et al., 2020).

* 1. **Distributed Computing for Churn Prediction**

Apache Spark has become essential for scaling machine learning workflows as there’s increasing size of datasets in modern applications. PySpark, the Python API for Spark, enables efficient processing of large datasets by distributing computations across multiple nodes. This capability enables PySpark particularly suitable for churn prediction tasks that involve analyzing large volume of customer data (Zaharia et al., 2016).

* 1. **Previous Work on Orange Telecom Dataset**

The Orange Telecom churn dataset has been widely used as benchmark for evaluating churn prediction models. Prior studies have applied various machine learning algorithms to this dataset, achieving notable results. Hadden et al. (2007) demonstrated the effectiveness of decision trees and logistic regression in identifying churn-prone customers, while more recent work has explored ensemble techniques to improve model accuracy (Burez & Van den Poel, 2009; Verbeke et al., 2012). However, few studies used PySpark to handle computational challenges posed by this dataset, leaving gap for this project.

By building on this body of work, the project develops predictive models for customer churn in addition to exploring application of PySpark for data analysis and machine learning.

1. **Methodology**

Spark ML and Spark MLlib packages were utilized, take advantage of their scalability and flexibility for handling large datasets. The methodology includes data processing, exploratory data analysis, feature engineering, model training, and evaluation, with detail below.

* 1. **Data Preprocessing**

The churn-80 dataset was processed and prepared for analysis using PySpark. The key processing tasks included:

* Handling Missing Values: missing data was addressed by removing incomplete rows.
* Categorical Variable Encoding: Features such as “International Plan” and “Voice Mail Plan” were converted into number values using Spark ML’s StringIndexer and OneHotEncoder to prepare for machine learning models.
* Data Normalization: Numerical features such as call durations and charges, were normalized using Spark ML’s StandardScaler to ensure that no feature disproportionately influenced model training.
  1. **Exploratory Data Analysis (EDA)**

EDA used to understand dataset’s structure, identify key trends, and detect anomalies. Data visualizations libraries such as Matplotlib were used to:

* Analyze the distribution of numerical variables such as “Total Day Minutes” and “Total Night Calls.”
* Explore correlations between features to avoid multicollinearity (e.g., “Total Day Minutes” vs. “Total Day Charges”).
* Assess churn rates across customer groups, such as those with or without international plan.
  1. **Feature Engineering**

Featured engineer was used for improving model performance. The steps included:

* Create New Features: “Total Charges” (aggregating day, evening and night charges) and “Average Call Duration” were introduced.
* Feature Selection: Tree-based models in Spark MLlib and correlation analysis helped identify the most predictive features.
  1. **Model Training**

Spark ML and Spark MLlib were used for model training and evaluation. The process involved:

* Pipeline: pipeline was created using Spark ML, comprising stages such as VectorAssember, StandardScaler, and a classifier. Pipelines streamlined with preprocessing, training and evaluation workflow.
* Algorithm Selection: several algorithms were tested including:
  + Logistic Regression: baseline model to reevaluate linear relationships.
  + Random Forests and Gradient-Boosted Trees: Tree-based models excel at capturing non-linear interactions.
  + Decision Trees: useful for understanding feature splits and interactions.
  + Support Vector Machines (SVMs): alternative for handling high-dimensional data.
* Train-Test Split: churn-80 dataset was split into training (80%) and validation (20%) sets. The churn-20 dataset was reserved for final testing.
* **Train-Test Split:** The churn-80 dataset was split into training (80%) and validation (20%) sets. The churn-20 dataset was reserved for final testing.
  1. **Hyperparameter Tuning**

Grid search and cross-validation were employed to optimized hyperparameters such as:

* Number of trees in random forests and gradient-boosted trees.
* Maximum depth and minimum split size for decision trees.
* Regularization parameters for logistic regression.
  1. **Model Evaluation**

The evaluation process focused on assessing model performance using key metrics:

* Accuracy: measure overall predictive performance.
* Precision and Recall: balance false positives and false negatives.
* F1-Score: provide a single metric for model effectiveness.
* ROC-AUC: assess model’s ability to distinguish between churned and non-churned customers.   
  1. **Deployment and Visualization**

The trained model was applied to the churn-20 datasets. By integrating both Spark ML and Spark MLlib capabilities, the project ensure efficient and scalable approach to solving the churn prediction problem, while enable flexibility in feature engineering, pipeline management, and algorithm selection.

1. **Experiment Design and Results**
   1. **Experiment Design**

The experiment was structured to evaluate the performance of machine learning models in predicting customer churn using the churn-80 dataset for training and cross-validation, and the churn-20 dataset for final testing. The primary objective was to identify the model that best balances accuracy, precision, recall, and overall interpretability for actionable business insights.

To ensure reliability and robustness, the following steps were implemented:

1. **Train-Test Split:** The churn-80 dataset was divided into 80% training and 20% validation subsets, while the churn-20 dataset served as an unseen test set.
2. **Cross-Validation:** A 5-fold cross-validation approach was employed during hyperparameter tuning to prevent overfitting.
3. **Evaluation Metrics:** Models were evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, to provide a comprehensive assessment of performance.
4. **Comparison of Algorithms:** Various algorithms, including Logistic Regression, Decision Trees, Random Forest, Gradient-Boosted Trees, and Support Vector Machines (SVM), were implemented to compare their strengths and weaknesses in predicting churn.  
   1. **Results**

The results of the experiment are summarized below, focusing on both model performance and insights derived from the predictions.

* + 1. **Model Performance**

After testing multiple algorithms, Random Forest and Gradient-Boosted Trees emerged as the top-performing models. The results for key metrics are as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** |
| Logistic Regression | 82.5% | 78.3% | 70.2% | 74.0% | 85.4% |
| Decision Tree | 88.1% | 84.7% | 79.3% | 81.9% | 90.2% |
| Random Forest | **91.3%** | **87.5%** | 85.4% | **86.4%** | **94.7%** |
| Gradient-Boosted Trees | 90.9% | 86.9% | **86.1%** | 86.5% | 94.3% |
| Support Vector Machines | 81.2% | 75.6% | 68.7% | 72.0% | 84.9% |

* **Random Forest** achieved the highest accuracy and ROC-AUC, making it the most reliable model for predicting churn.
* **Gradient-Boosted Trees** provided competitive performance with slightly higher recall, which could be advantageous for minimizing false negatives (i.e., retaining at-risk customers).  
  + 1. **Insights from Feature Importance**

Using the feature importance scores provided by tree-based models, the following features were identified as the most predictive of customer churn:

* **Customer Service Calls:** Higher numbers of calls to customer service were strongly correlated with churn, likely indicating dissatisfaction.
* **Total Day Minutes:** Customers with higher daytime call durations were more likely to churn, possibly reflecting heavy users evaluating alternatives.
* **International Plan:** Customers enrolled in the international plan showed lower churn rates, suggesting higher satisfaction or loyalty among these customers.
* **Total Charges:** Higher total charges, combining day, evening, and night charges, were indicative of churn, highlighting the potential impact of pricing.  
  + 1. **Validation on Test Set**

The final evaluation on the churn-20 test dataset confirmed the robustness of the Random Forest model, with results closely matching the validation metrics. The model achieved:

* **Accuracy:** 91.0%
* **Precision:** 87.2%
* **Recall:** 84.9%
* **F1-Score:** 86.0%
* **ROC-AUC:** 94.5%

1. **Summary and Future Work**

Customer churn predication is a very important task for modern business such as telecom industry operate in highly competitive markets where retaining customer is more cost-effective than acquiring new ones. In this project, Apache Spark’s MLlib and ML packages to build a robust and scalable pipeline for predicting customer churn. The analysis provides actionable insights into customer behavior and identify key predictors of churn such as customer service interactions, usage patterns and subscription plans.

Through experimentation with multiple machine learning models, Random Forest and Gradient-Boosted Trees are top performers, achieving high accuracy, precision, recall and ROC-AUC.

While the results of this project are promising, several areas for future exploration and improvement remain:

1. Incorporate additional features: Expand dataset with more attributes such as billing history, customer demographics, and social media sentiment.
2. Explore Advanced Models: Deep learning models such as Recurrent Neural Networks (RNNs) or Transformer-based architectures, could be explored for their ability to capture patterns over time.
3. Real-Time Predictions: By using streaming data and integrating with customer relationship management (CRM) systems could enable immediate and personalized retention strategies.

By implementing these ideas, the predictive models in this project could evolve into a comprehensive and impact solution. The future lies in combing data-driven insights with real-time operational capabilities to drive customer loyalty and business success.

**References**

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36(3),* 4626-4636. <https://doi.org/10.1016/j.eswa.2008.05.027>

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, 34(10), 2902-2917. https://doi.org/10.1016/j.cor.2005.11.007

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2), 2592-2602. https://doi.org/10.1016/j.eswa.2008.02.021

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit-driven data mining approach. *European Journal of Operational Research, 218(1),* 211-229. https://doi.org/10.1016/j.ejor.2011.09.031

Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., & Stoica, I. (2016). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (NSDI)*.