1. Project Title

Customer Churn Prediction Using Classification Algorithms and Survival Analysis

2. Abstract

Customer churn, the loss of clients or subscribers, is a critical issue for many businesses, particularly those in the subscription-based service industry. This project aims to develop a predictive model that can identify customers who are likely to leave a service or cancel their subscription. Using a combination of customer interaction logs, transaction history, and demographic data, we will apply classification algorithms such as decision trees and logistic regression, as well as survival analysis, to predict churn. The outcomes of this project will help businesses proactively address churn, improving customer retention and overall profitability.

3. Introduction

In highly competitive markets, retaining existing customers is often more cost-effective than acquiring new ones. Customer churn refers to the phenomenon where customers stop using a service or discontinue their subscriptions. Understanding and predicting which customers are likely to churn can provide companies with valuable insights to take preventive actions. The objective of this project is to develop a predictive model that can accurately forecast customer churn, enabling businesses to intervene and reduce churn rates.

This project will utilize various data sources such as customer interaction logs, transaction history, and demographic data to build a comprehensive picture of customer behavior. By leveraging advanced data mining techniques, including classification algorithms and survival analysis, the project aims to create a robust model capable of predicting churn with high accuracy. The predictive model will be instrumental for businesses looking to enhance customer loyalty and optimize their marketing strategies.

4. Related Work

Predicting customer churn has been a topic of extensive research in data mining and customer relationship management (CRM). Various studies have employed machine learning techniques to tackle this problem, with logistic regression and decision trees being among the most popular methods due to their interpretability and effectiveness.

For instance, in a study by Verbeke et al. (2012), the authors compared several classification techniques for churn prediction and found that decision trees, when combined with ensemble methods, provided high predictive performance. Another study by Neslin et al. (2006) focused on the application of logistic regression in churn prediction and demonstrated its effectiveness when integrated with behavioral data.

Survival analysis, traditionally used in medical research, has also been adapted for churn prediction. In the work by Buckinx and Van den Poel (2005), survival analysis was employed to model customer retention and identify factors that influence churn over time. This approach is particularly useful as it accounts for the time dimension, providing insights into not just whether a customer will churn, but also when they are likely to do so.

While these studies provide a solid foundation, there is still a need for more research that combines multiple data sources and techniques to improve predictive accuracy. This project intends to build on these prior works by integrating demographic data with behavioral logs and applying both classification and survival analysis techniques.

5. Proposed Work and Progress

Data Collection:

- **Completed Tasks:**
  - **Data was collected from customer interaction logs, transaction history, and demographic data, utilizing the IBM Watson Telco Customer Churn dataset.**
  - **Preliminary data exploration and consolidation were successfully completed.**

Data Preprocessing:

- **Completed Tasks:**
  - **Data cleaning: Addressed missing values, outlier detection, and normalization.**
  - **Feature engineering: Created features such as customer tenure, interaction frequency, and purchase patterns.**
  - **Data splitting: The dataset was split into 70% training and 30% testing sets.**

Model Development:

- **Completed Tasks:**
  - Developed classification models using decision trees and logistic regression.
  - Implemented survival analysis techniques, including Kaplan-Meier estimator and Cox Proportional Hazards model.
  - Hyperparameter tuning and initial model evaluations using cross-validation**.**

Challenges Faced:

- **Data Imbalance:** The churn dataset was highly imbalanced, with a small percentage of customers classified as churners. Addressed using techniques like SMOTE (Synthetic Minority Over-sampling Technique).
- **Feature Correlation:** High correlation between some features led to multicollinearity issues, requiring careful feature selection and regularization techniques.

Current Phase: Model Evaluation and Iteration (Weeks 11-13)

6. Evaluation

Performance Metrics:

- **Completed Evaluations:**
    - **Classification models:** Achieved an accuracy of 85%, precision of 78%, recall of 65%, F1-score of 71%, and AUC-ROC of 0.83.
    - **Survival analysis:** C-index of 0.76, indicating good predictive power for the timing of churn.

Cross-Validation:

- **Completed Tasks:**
    - K-fold cross-validation (k=10) was used to validate model performance, ensuring robustness and preventing overfitting**.**

Baseline Comparison:

- **Completed Comparisons:**
    - Baseline models (simple decision tree, logistic regression without feature engineering) achieved lower performance, with AUC-ROC scores around 0.75, highlighting the improvement from advanced techniques.

Business Impact Evaluation:

- **Ongoing Tasks:**
    - Conducting a cost-benefit analysis to weigh retention strategy costs against potential revenue saved by preventing churn.
    - Analyzing the confusion matrix to understand the types of errors and their business implications.

Challenges Faced:

- **Metric Trade-offs:** Balancing precision and recall has been challenging, as improving one often reduces the other. Focused on optimizing the F1-score for a balanced approach.

7. Discussion

Project Timeline Update:

1. **Phase 1: Project Planning and Data Collection (Weeks 1-3)**
   - Completed as planned.
2. **Phase 2: Data Preprocessing and Feature Engineering (Weeks 4-6)**
   - Completed as planned.
3. **Phase 3: Model Development (Weeks 7-10)**
   - Completed as planned.
4. **Phase 4: Model Evaluation and Iteration (Weeks 11-13)**
   - **Week 11-12**: Evaluated models against success criteria. Identified the logistic regression model with feature engineering as the best-performing model.
   - **Week 13**: Iterating on the model to improve recall without significantly compromising precision. Fine-tuning the survival analysis model for better accuracy in predicting churn timing.

Implementation and Deployment (Upcoming):

- **Weeks 14-15**: Plan to implement the final model in a production environment, integrating it with business processes for real-time churn prediction. Developing a user-friendly dashboard for visualizing predictions and key metrics.
- **Week 16**: Finalize documentation, including a detailed report on the project methodology, results, and business implications. Conduct a presentation or workshop to demonstrate the model's capabilities to stakeholders.

8. Conclusion

Customer churn poses a significant challenge for businesses, particularly in subscription-based industries where retaining existing customers is crucial for sustaining revenue growth. This project aims to address this challenge by developing a predictive model that accurately forecasts customer churn using a combination of classification algorithms and survival analysis techniques.

By leveraging diverse data sources such as customer interaction logs, transaction history, and demographic information, the project seeks to create a comprehensive and actionable churn prediction model. The evaluation plan emphasizes not only predictive accuracy but also the practical impact of the model on business outcomes, such as reducing churn rates and increasing customer lifetime value.

The proposed timeline for the project is designed to ensure thorough data preparation, model development, and evaluation, culminating in the deployment of a robust model that can be integrated into business operations. The success of this project will be measured by the model's ability to provide actionable insights and deliver tangible business value, ultimately helping companies enhance customer retention and achieve long-term growth.

The completion of this project will provide businesses with a powerful tool to predict and mitigate customer churn, enabling more proactive and targeted retention strategies.