Customer Churn Prediction Using Classification Algorithms and Survival Analysis

FirstName Surname
Department Name
Institution/University Name
City State Country
email@email.com

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.

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.

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.

PROPOSED WORK

The proposed project will involve several key tasks, as outlined below:

1. Data Collection:

Data Sources: The project will utilize customer interaction logs (e.g., customer service calls, website visits), transaction history (e.g., purchase records, subscription renewals), and demographic data (e.g., age, location, income level).

Data Acquisition: Data will be collected from company databases or publicly available datasets such as the IBM Watson Telco Customer Churn dataset.

2. Data Preprocessing:

Cleaning: Handling missing values, outlier detection, and data normalization.

Feature Engineering: Creating new features such as customer tenure, frequency of interaction, and purchase patterns.

Data Splitting: The dataset will be split into training and testing sets to evaluate model performance.

3. Model Development:

Classification Algorithms:

- **Decision Trees**: A non-parametric method that splits the data into subsets based on feature values to predict the target variable (churn/no churn).
- **Logistic Regression**: A statistical model that predicts the probability of churn based on independent variables (customer features).
- **Survival Analysis**: Techniques like Kaplan-Meier estimator and Cox Proportional Hazards model will be used to estimate the time until churn occurs.

4. Model Evaluation:

Metrics: Performance will be assessed using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) for classification models. For survival analysis, concordance index (C-index) will be used.

Cross-Validation: K-fold cross-validation will be employed to ensure the robustness of the models.

5. Implementation and Deployment:

The final model will be implemented using Python libraries such as Scikit-learn, Lifelines, and Pandas.

A dashboard will be created for businesses to visualize churn predictions and key customer metrics.

6. EVALUATION

The evaluation of the customer churn prediction model will focus on both the predictive accuracy and the practical utility of the model. The following plan outlines the key aspects of the evaluation process:

6.1 Performance Metrics:

Accuracy: Measures the overall correctness of the model in predicting churn versus non-churn. However, given the typically imbalanced nature of churn data, this metric alone may not fully capture model performance.

Precision and Recall: Precision will indicate the proportion of predicted churns that are actual churns, while recall (or sensitivity) will show the proportion of actual churns that were correctly identified by the model. These metrics are particularly important when false positives and false negatives carry different business implications.

F1-Score: Combines precision and recall into a single metric, providing a balance between the two.

ROC-AUC (Receiver Operating Characteristic - Area Under Curve): This metric provides a measure of the model's ability to distinguish between churners and non-churners, regardless of the threshold chosen.

C-Index (Concordance Index) for Survival Analysis: This metric will evaluate the accuracy of the survival analysis in predicting the time until churn. A higher C-Index indicates a better model.

6.2 Cross-Validation:

K-Fold Cross-Validation: The model will be validated using k-fold cross-validation, typically with k=5 or k=10, to ensure that the model's performance is consistent across different subsets of the data. This technique helps prevent overfitting and gives a more reliable estimate of the model's predictive capability.

6.3 Baseline Comparison:

Baseline Models: The performance of the proposed model will be compared against baseline models, such as a simple decision tree, random guessing, or a logistic regression without engineered features. This comparison will help demonstrate the added value of the more sophisticated techniques employed.

6.4 Business Impact Evaluation:

Cost-Benefit Analysis: Beyond technical performance, the model's effectiveness will be measured by its impact on business metrics. This includes a cost-benefit analysis, where the costs of implementing retention strategies for predicted churners are weighed against the potential revenue saved by retaining these customers.

Confusion Matrix Analysis: Understanding the types of errors the model makes (e.g., false positives vs. false negatives) will provide insights into how the model's predictions align with business goals.

6.5 Success Criteria:

High Predictive Accuracy: A successful model should achieve a high ROC-AUC score (e.g., above 0.8), indicating strong discrimination between churners and non-churners.

Actionable Insights: The model should not only predict churn but also provide interpretable results that can guide business strategies. For

example, identifying key predictors of churn that can be targeted for customer retention efforts.

Business Value: Ultimately, success will be measured by the model's ability to reduce churn rates and increase customer lifetime value (CLTV), providing a clear return on investment (ROI) for the company.

7. DISCUSSION

The proposed project is planned to be completed over a span of approximately four months. Below is a detailed timeline outlining the phases of the project and their respective durations:

Phase 1: Project Planning and Data Collection (Weeks 1-3)

Week 1: Finalize the project scope, objectives, and deliverables. Identify and secure access to the required datasets.

Weeks 2-3: Collect and consolidate data from various sources (e.g., interaction logs, transaction history, demographic data). Begin preliminary data exploration to understand the structure and characteristics of the data.

Phase 2: Data Preprocessing and Feature Engineering (Weeks 4-6)

Week 4: Perform data cleaning, including handling missing values, outliers, and normalization.

Weeks 5-6: Conduct feature engineering to create new variables that could enhance the model's predictive power (e.g., customer tenure, frequency of transactions). Split the data into training and testing sets.

Phase 3: Model Development (Weeks 7-10)

Weeks 7-8: Develop classification models using decision trees and logistic regression. Tune hyperparameters and evaluate models using cross-validation.

Weeks 9-10: Implement survival analysis techniques to predict the time until churn. Evaluate the performance of the survival models using the C-Index.

Phase 4: Model Evaluation and Iteration (Weeks 11-13)

Weeks 11-12: Evaluate the models against the success criteria outlined in the evaluation plan. Compare the performance of different models and select the best-performing model.

Week 13: Iterate on the model, making improvements based on evaluation feedback. Fine-tune the model to enhance predictive accuracy and interpretability.

Phase 5: Implementation and Deployment (Weeks 14-16)

Weeks 14-15: Implement the final model in a production environment, integrating it with business processes for real-time churn prediction. Develop 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.

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.

REFERENCES