



CUSTOMER CHURN PREDICTION USING CLASSIFICATION ALGORITHMS AND SURVIVAL ANALYSIS

August 2024

Customer Churn Prediction Using Classification Algorithms and Survival Analysis

- **Critical Issue:** Customer churn, the loss of clients or subscribers, poses significant challenges for businesses, particularly in subscription-based industries.
- **Project Aim:** Develop a predictive model utilizing classification algorithms and survival analysis to identify customers likely to churn, enabling proactive retention strategies.
- **Data Utilized:** Model integrates customer interaction logs, transaction history, and demographic data to enhance predictive accuracy.
- **Outcome:** The model is expected to improve customer retention and overall profitability for businesses by providing actionable insights.


Introduction

- **Customer Churn Definition:** Customer churn occurs when clients stop using a service or cancel their subscriptions, impacting business profitability.
- **Project Objective:** Develop a predictive model to forecast customer churn, enabling businesses to take preventive actions and reduce churn rates.
- **Data Sources:** Utilizes customer interaction logs, transaction history, and demographic data to build a predictive model.
- **Significance:** Accurate churn prediction helps businesses improve customer retention, enhancing loyalty and optimizing marketing strategies.

Related Work

- **Machine Learning Techniques:** Logistic regression and decision trees are commonly used due to their interpretability and effectiveness in predicting churn.
- **Survival Analysis:** Adapted from medical research, survival analysis models customer retention and identifies factors influencing churn over time.
- **Research Gaps:** Prior studies focus on individual techniques; this project integrates multiple data sources and combines classification with survival analysis.

Proposed Work

- **Data Collection:** Utilize customer interaction logs, transaction history, and demographic data. Sources include company databases and publicly available datasets.
 - **Data Preprocessing:** Includes data cleaning, feature engineering, and data splitting for model training and testing.
 - **Model Development:** Apply classification algorithms (decision trees, logistic regression) and survival analysis to predict churn.
 - **Model Evaluation:** Assess model performance using accuracy, precision, recall, F1-score, AUC-ROC, and C-index for survival analysis.
- 


Data Collection Methods

- **Data Sources:** Utilize customer interaction logs, transaction history, and demographic data to build the predictive model.
 - **Data Acquisition:** Collect data from company databases and publicly available datasets like IBM Watson Telco Customer Churn dataset.
 - **Completed Tasks:** Successfully collected and consolidated data, ready for preprocessing and model development.
-

Data Preprocessing Techniques

- **Data Cleaning:** Addressed missing values, outlier detection, and normalization to prepare the data for modeling.
- **Feature Engineering:** Created new features such as customer tenure, interaction frequency, and purchase patterns.
- **Data Splitting:** Split the dataset into 70% training and 30% testing sets to evaluate model performance.


Feature Engineering

- **Creating New Features:** Developed new features such as customer tenure, frequency of interaction, and purchase patterns to improve model accuracy.
 - **Handling Correlations:** Addressed multicollinearity issues by selecting and regularizing features with high correlations.
 - **Feature Importance:** Identified key features that significantly impact churn prediction, aiding in model interpretability.
- 

Classification Algorithms

- **Decision Trees:** A non-parametric method that splits the data into subsets based on feature values to predict customer churn.
- **Logistic Regression:** A statistical model that predicts the probability of churn based on independent variables such as customer features.
- **Model Tuning:** Hyperparameter tuning and cross-validation were used to optimize model performance and prevent overfitting.

Survival Analysis

- **Purpose:** Survival analysis models the time until churn occurs, providing insights into when customers are likely to churn.
 - **Techniques Used:** Kaplan-Meier estimator and Cox Proportional Hazards model were applied to estimate and analyze the time-to-event data.
 - **Applications:** Useful for understanding customer lifecycle and timing retention efforts to maximize effectiveness.
- 

Model Evaluation Metrics

- **Classification Metrics:** Performance assessed using accuracy, precision, recall, F1-score, and AUC-ROC for model evaluation.
- **Survival Analysis Metric:** Concordance Index (C-index) was used to evaluate the predictive power of survival analysis models.
- **Cross-Validation:** K-fold cross-validation employed to ensure robustness and prevent overfitting of the models.

Implementation and Deployment

- **Implementation Tools:** The model was implemented using Python libraries such as Scikit-learn, Lifelines, and Pandas.
- **Dashboard Development:** A dashboard was created for businesses to visualize churn predictions and key customer metrics.
- **Deployment Strategy:** The model was deployed in a production environment to facilitate real-time decision-making.

Model Performance Results

- **Classification Models:** Achieved 85% accuracy, 78% precision, 65% recall, 71% F1-score, and AUC-ROC of 0.83.
- **Survival Analysis:** C-index of 0.76, indicating good predictive power for estimating the timing of customer churn.
- **Comparison to Baseline:** Advanced techniques significantly outperformed baseline models, such as simple decision trees and logistic regression without feature engineering.

Business Impact Evaluation

- **Cost-Benefit Analysis:** Evaluated the costs of implementing retention strategies against the potential revenue saved by retaining customers.
- **Confusion Matrix Analysis:** Analyzed errors to understand their business implications, focusing on false positives and negatives.
- **Business Success Criteria:** Success measured by the model's ability to reduce churn rates, increase customer lifetime value, and provide a return on investment.

Challenges and Solutions

- **Data Imbalance:** Addressed using techniques like SMOTE (Synthetic Minority Over-sampling Technique) to handle the imbalance in churn vs. non-churn classes.
 - **Metric Trade-offs:** Balanced precision and recall by optimizing the F1-score to achieve a balanced model performance.
 - **Feature Correlation:** Managed multicollinearity by careful feature selection and regularization techniques to ensure model stability.
-


Key Findings

- **Effective Models:** Logistic regression with feature engineering provided the best balance of precision and recall, with strong AUC-ROC scores.
- **Survival Analysis Insights:** Provided valuable insights into the timing of churn, aiding businesses in prioritizing retention efforts.
- **Business Impact:** Model showed potential to reduce churn rates and increase customer lifetime value, offering a strong return on investment (ROI).

Conclusion

- **Project Success:** Successfully developed a predictive model using classification algorithms and survival analysis, achieving high accuracy and actionable insights.
 - **Business Impact:** The model provides businesses with a powerful tool to reduce churn, increase customer lifetime value, and enhance profitability.
 - **Future Improvements:** Further work could explore advanced techniques like deep learning, incorporate additional data sources, and implement real-time updates.
-

Future Work

- **Advanced Techniques:** Explore deep learning models to potentially improve predictive accuracy further.
 - **Additional Data Sources:** Incorporate data from social media, customer feedback, and sentiment analysis to enrich the model's predictive capabilities.
 - **Real-Time Updates:** Implement real-time data updates to enhance the model's responsiveness to changing customer behavior.
- 
- A solid green horizontal bar spanning the width of the slide, located at the bottom.

References

- **Key Sources:** Verbeke et al. (2012) on classification techniques, Neslin et al. (2006) on logistic regression, and Buckinx and Van den Poel (2005) on survival analysis in churn prediction.
- **Additional Literature:** Further studies and research articles referenced in the project, focusing on advanced data mining and machine learning techniques.

Thank You & Q&A

- **Thank You for Your Attention:** We appreciate your time and attention during this presentation.
- **Q&A Session:** Please feel free to ask any questions or share your thoughts.