**Lauren Butcher**  
DSC 680- White Paper  
July 13, 2024

Customer Churn Prediction in E-commerce

**Business Problem**

Reducing customer churn rate is critical for e-commerce businesses to increase revenue and retain customers. By predicting which customers are likely to churn, the business can implement targeted retention strategies to improve customer loyalty and reduce revenue loss.

**Background/History**

Customer churn prediction has become a vital aspect of business strategy for e-commerce companies. High churn rates can significantly impact a company's profitability and market position. Traditional methods focused on historical data analysis, but advancements in machine learning have enabled more accurate and actionable predictions.

**Data Explanation**

**Datasets**:

1. **Customer Churn Prediction on Kaggle**:
   * **Description**: This dataset provides various customer information, including demographics, account details, and service usage, to predict churn.
   * **URL**: [Customer Churn Prediction on Kaggle](https://www.kaggle.com/code/bhartiprasad17/customer-churn-prediction)
2. **Ecommerce Customer Churn Analysis and Prediction on Kaggle**:
   * **Description**: This dataset includes detailed information about customer demographics and online behavior, designed to help predict customer churn.
   * **URL**: [Ecommerce Customer Churn Analysis and Prediction](https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction)
3. **E-Commerce Customer Churn Analysis and Prediction on Kaggle**:
   * **Description**: Another comprehensive dataset focusing on customer churn analysis in e-commerce, including various features for building predictive models.
   * **URL**: [E-Commerce Customer Churn Analysis and Prediction](https://www.kaggle.com/code/ronaldopangarego/e-commerce-customer-churn-analysis-and-prediction)
4. **E-commerce Customer Churn Prediction on Kaggle**:
   * **Description**: This dataset offers data for predicting churn in an e-commerce setting, useful for building and evaluating different predictive models.
   * **URL**: [E-commerce Customer Churn Prediction](https://www.kaggle.com/code/wonderdavid/e-commerce-customer-churn-prediction)

The datasets include features such as customer demographics, transaction history, service usage patterns, and more. Data preparation involved handling missing values, normalizing data, and feature engineering to create meaningful variables for analysis.

**Methods**

To analyze and predict customer churn, the following methods will be used:

* Logistic Regression
* Random Forest
* Neural Networks

These methods allow for building robust predictive models and understanding the key factors contributing to customer churn.

**Analysis**

The analysis involves training and validating different machine learning models on the provided datasets. Techniques such as cross-validation and grid search are used to optimize model performance. The imbalanced nature of the datasets requires special handling, such as oversampling the minority class or using cost-sensitive learning methods.

**Conclusion**

Predicting customer churn in e-commerce is essential for implementing effective retention strategies. The models developed in this project provide actionable insights into customer behavior, enabling businesses to proactively address churn risks.

**Assumptions**

* The data provided is representative of the overall customer base.
* The features selected for the models are sufficient to predict churn accurately.

**Limitations**

* The models may not generalize well to different e-commerce platforms with varying customer behaviors.
* Data quality issues, such as missing or inaccurate entries, can affect model performance.

**Challenges**

* Handling the imbalanced dataset to ensure accurate predictions.
* Feature engineering to create meaningful variables from raw data.
* Selecting appropriate metrics and validation strategies for model evaluation.

**Future Uses/Additional Applications**

* Extending the churn prediction model to other industries, such as telecommunications or banking.
* Integrating the model with real-time data processing systems to provide instant churn predictions.
* Using the insights from churn prediction to develop personalized marketing campaigns.

**Recommendations**

* Regularly update the model with new data to maintain its accuracy.
* Implement targeted retention strategies based on the model's predictions.
* Continuously monitor model performance and adjust features and algorithms as needed.

**Implementation Plan**

1. **Data Collection**: Gather and preprocess data from various sources.
2. **Model Training**: Train machine learning models using the prepared datasets.
3. **Validation**: Validate the models using appropriate metrics and validation strategies.
4. **Deployment**: Deploy the best-performing model in a production environment.
5. **Monitoring**: Continuously monitor model performance and make necessary adjustments.

**Ethical Assessment**

* **Data Privacy**: Ensuring the protection of customer data and complying with data privacy regulations.
* **Avoiding Discrimination**: Making sure the model does not discriminate against any group of customers based on protected attributes like age, gender, or ethnicity.

**Visualizations**

1. **Churn Rate by Gender** (Telco Dataset)
   * A graph of a number of people

     Description automatically generated
2. **Churn Rate by Internet Service Type** (Telco Dataset)
   * A graph of different colored bars

     Description automatically generated
3. **Monthly Charges Distribution for Churned and Non-Churned Customers** (Telco Dataset)
   * A graph showing a number of charges

     Description automatically generated
4. **Tenure vs. Monthly Charges Scatter Plot** (Telco Dataset)
   * A graph showing a number of different colored crosses

     Description automatically generated with medium confidence
5. **Churn Rate by Preferred Login Device** (E-Commerce Dataset)
   * A graph of a login device

     Description automatically generated

**Appendix**

**Supporting Documentation**: Detailed data dictionary, model training logs, and code snippets used for data preprocessing and model training.

**10 Questions an Audience Would Ask**

1. **How was the data collected and processed?**

* The data was collected from publicly available Kaggle datasets. The data includes customer demographics, account details, and service usage patterns. Data processing involved handling missing values, normalizing data, and feature engineering to create meaningful variables for analysis.

1. **What features were most important in predicting churn?**

* **Using the Random Forest model, the most important features in predicting churn include:**
  + Tenure
  + MonthlyCharges
  + Contract
  + TotalCharges
  + InternetService
  + PaymentMethod
  + TechSupport
  + OnlineSecurity
  + Dependents
  + PhoneService

1. **How did you handle missing data and outliers?**

* Missing data was handled by filling missing values with appropriate methods such as mean, median, or mode for numerical columns and the most frequent value for categorical columns. Outliers were detected and handled using techniques like removing or capping the outlier values.

1. **What metrics were used to evaluate model performance?**

* The metrics used to evaluate model performance include accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provide a comprehensive evaluation of the model's predictive power and its ability to handle imbalanced datasets.

1. **How do you ensure the model is not biased?**

* To ensure the model is not biased, we implemented techniques such as stratified sampling during train-test split, feature importance analysis to ensure no discriminatory features are included, and regular monitoring of model predictions to check for biases.

1. **What are the potential impacts of incorrect predictions?**

* Incorrect predictions can lead to loss of customers who might have been retained with appropriate interventions. Additionally, false positives (predicting a customer will churn when they won't) can lead to unnecessary retention efforts and costs.

1. **How frequently should the model be retrained?**

* The model should be retrained periodically, such as monthly or quarterly, to incorporate new data and adapt to changing customer behavior patterns. Continuous monitoring and evaluation of model performance will help determine the optimal retraining frequency.

1. **How do you integrate the model with existing business systems?**

* The model can be integrated with existing business systems through APIs or batch processing systems. This allows the model to score customers in real-time or on a scheduled basis and provide actionable insights to customer retention teams.

1. **What are the costs associated with implementing the model?**

* The costs include data storage and processing, computational resources for model training and deployment, and costs associated with integrating the model into business systems. Additionally, there may be costs for ongoing monitoring and maintenance of the model.

1. **How do you measure the success of the retention strategies based on the model's predictions?**

* The success of retention strategies can be measured by tracking key metrics such as churn rate reduction, customer lifetime value, customer satisfaction scores, and the ROI of retention campaigns. Comparing these metrics before and after implementing the model-based strategies will help measure success.

**References**:

* [Customer Churn Prediction on Kaggle](https://www.kaggle.com/code/bhartiprasad17/customer-churn-prediction)
* [Ecommerce Customer Churn Analysis and Prediction on Kaggle](https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction)
* [E-Commerce Customer Churn Analysis and Prediction on Kaggle](https://www.kaggle.com/code/ronaldopangarego/e-commerce-customer-churn-analysis-and-prediction)
* [E-commerce Customer Churn Prediction on Kaggle](https://www.kaggle.com/code/wonderdavid/e-commerce-customer-churn-prediction)
* Relevant research papers and articles on customer churn prediction and machine learning techniques.