**Customer Churn Prediction Analysis**

**Submitted for**

**CSET211 - Statistical Machine Learning**

Submitted by:

**(E23CSEU0014) ABHIRAJ GHOSE**

**(E23CSEU0022) PALLAV SHARMA**

**(E23CSEU0112) VAIBHAV GUPTA**

A close-up of a logo

Description automatically generated

Submitted to:

**Mr Prashant Kapil**

**July-Dec 2024**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

**INDEX**

**Abstract**

1. **Introduction**
2. **Related Survey**
3. **Datasets**

**3.1 Data Preprocessing**

1. **Methodology**
   1. **Hardware and Software Requirements**
   2. **Performance Metrics**
2. **Results and Analysis**
3. **Conclusions and Future Works**

**\*\* Include the GitHub link as the last sentence of the Abstract for easy access to supplementary materials.**

**Abstract**

This project focuses on predicting customer churn using machine learning techniques. The objective was to analyze patterns in customer data and build models to classify churn behaviour effectively. Key methods included data preprocessing, feature engineering, and implementing Logistic Regression and Random Forest Classifier. The models were evaluated using metrics like accuracy, precision, recall, F1 score, and ROC-AUC curve. The results demonstrated significant insights into churn drivers and provided actionable recommendations for customer retention strategies.  
**GitHub Repository**: [Insert GitHub link here]

**1. Introduction**

This project addresses the issue of customer churn prediction, which is critical for businesses aiming to retain customers and reduce revenue loss. Using the publicly available Telco Customer Churn dataset, the project's objectives include developing accurate predictive models, identifying key factors contributing to churn, and providing actionable insights to enhance customer retention strategies.

**Key Contributions:**

* Preprocessing and feature engineering for robust model training.
* Comparison of multiple classifiers to identify the best-performing model.
* Use of SHAP values to interpret model predictions.

**2. Related Survey**

Several studies have utilized machine learning to predict customer churn. However, gaps remain in interpretability and real-world application. This project bridges these gaps by combining interpretable models like Logistic Regression with ensemble methods and SHAP for feature importance analysis, ensuring actionable insights.  
An in-depth report on Related Projects has been uploaded in Github.

**3. Datasets**

The project used the **Telco Customer Churn dataset** from Kaggle. Key features include customer demographics, account details, and service usage statistics.

**3.1 Data Preprocessing**

* **Missing Values**: Rows with missing data were dropped after imputation for numerical columns where feasible.
* **Feature Encoding**: One-hot encoding was applied to categorical variables.
* **Scaling**: Numerical features were scaled using StandardScaler.

**4. Methodology**

The workflow consisted of several key steps:

**4.1 Hardware and Software Requirements**

* **Hardware**: Laptop with high-performance specifications.
* **Software**: Python libraries including pandas, scikit-learn, shap, plotly, and seaborn.

**4.2 Performance Metrics**

* **Accuracy**: To measure the proportion of correctly predicted instances.
* **Precision, Recall, and F1 Score**: To evaluate the model's ability to balance false positives and false negatives.
* **ROC-AUC Curve**: To assess the classifier's ability to distinguish between classes.

**5. Results and Analysis**

**Logistic Regression**

* **Accuracy**: ~82%
* **Key Findings**: Provided baseline insights into feature importance and model interpretability.

**Random Forest**

* **Accuracy**: ~79% (after hyperparameter tuning).
* **Feature Importance**: Highlighted tenure, MonthlyCharges, and Contract as key factors.

**Analysis**

* **SHAP Interpretations**: SHAP values were used to understand feature impacts at both global and individual levels.
* **Correlation Analysis**: Features like Contract and InternetService showed strong associations with churn.

**6. Conclusions and Future Works**

**Achievements**

* Built and compared multiple predictive models for churn prediction.
* Provided actionable insights into customer retention strategies through data visualization and SHAP analysis.

**Future Directions**

* Incorporate deep learning models for potential performance improvement.
* Extend the dataset to include real-time data for better generalization.
* Explore text analytics to understand customer feedback for churn prediction.

Github: https://github.com/AetherSparks/Customer-Churn-Prediction