**Customer Churn Prediction Project Report**

**1. Introduction**

Customer churn is a critical problem for subscription-based businesses, as retaining existing customers is often more cost-effective than acquiring new ones. This project aims to develop a machine learning model to predict customer churn, enabling businesses to take proactive measures to retain customers.

**2. Dataset Overview**

The dataset consists of customer records with various attributes such as demographics, subscription details, and usage patterns. The target variable indicates whether a customer has churned (1) or remained subscribed (0).

* **Training Set:** customer\_churn\_dataset-training-master.csv (50,000 records)
* **Testing Set:** customer\_churn\_dataset-testing-master.csv (38,167 records)

**3. Exploratory Data Analysis (EDA)**

**Summary Statistics**

* Checked missing values and handled them appropriately.
* Examined distributions of numerical and categorical features.
* Identified correlations between features and the target variable.

**Data Visualization**

* Histograms and box plots for numerical features.
* Bar charts for categorical features.
* Correlation heatmap to understand relationships between variables.

**4. Data Preprocessing**

* **Handling Missing Values:** Imputed missing data where necessary.
* **Encoding Categorical Variables:** Used one-hot encoding for categorical features.
* **Feature Scaling:** Applied standardization for numerical features.
* **Class Imbalance Handling:** Used techniques like oversampling (SMOTE) or class-weight adjustments.

**5. Model Training & Evaluation**

We explored several classification algorithms to find the best-performing model.

**Algorithms Used**

* Logistic Regression
* Decision Tree
* Random Forest
* Support Vector Machine (SVC)
* CatBoost
* XGBoost

**Model Performance Metrics**

Since churn prediction is an imbalanced classification problem, we focused on precision, recall, and F1-score rather than just accuracy.

**Random Forest Results (Best Model)**

* **Accuracy:** 99.48%
* **Precision:** 99%
* **Recall:** 100%
* **F1-score:** 99%

**Cross-Validation Scores**

[0.99967, 0.99980, 0.99964, 0.99975, 0.99985]

Mean Accuracy: 99.97%

**6. Feature Importance**

The most significant features influencing customer churn were:

1. **Feature 5** – 31.8%
2. **Feature 8** – 21.5%
3. **Feature 1** – 13.5%
4. **Feature 6** – 12.4%

**7. Model Deployment**

The trained model was saved as a pickle file (customer\_churn\_model.pkl) for deployment and further analysis.

**8. Conclusion**

* The Random Forest model achieved high accuracy and recall, making it a reliable choice for predicting churn.
* Businesses can use these predictions to identify at-risk customers and implement targeted retention strategies.
* Future improvements could involve hyperparameter tuning, additional feature engineering, and testing more advanced deep learning models.