**Customer Churn Analysis Report**

**Introduction**

Customer churn is a critical challenge in the telecom industry, impacting revenue and customer retention. This project aims to predict churn for a telecom company using the Telco Customer Churn dataset (7,043 customers, 21 features). By leveraging machine learning, we identify key drivers of churn, segment customers, and provide actionable retention strategies. The project was executed in Visual Studio Code (VS Code), delivering a Python notebook, a PowerPoint presentation, and recommendations.

**Abstract**

This project predicts customer churn using a Random Forest Classifier, achieving an AUC-ROC of 0.82 and a recall of 0.47 for the churn class. We performed SQL-based data aggregation, exploratory data analysis (EDA), model explainability with ELI5 and SHAP, and customer segmentation with K-Means. Key findings include short tenure, high monthly charges, and fiber optic service as major churn drivers. Customers were segmented into At Risk, Loyal, and Dormant groups, enabling targeted retention strategies. Recommendations include offering long-term contract incentives and enhancing tech support to reduce churn.

**Tools Used**

* **Environment**: Visual Studio Code (VS Code) with Python extensions.
* **Programming Language**: Python 3.12.9.
* **Libraries**:
  + Data Handling: pandas, numpy, sqlite3.
  + Visualization: matplotlib, seaborn, SHAP.
  + Machine Learning: scikit-learn (Random Forest, K-Means).
  + Explainability: ELI5, SHAP.
* **Dataset**: Telco Customer Churn (Kaggle, 7,043 rows, 21 columns).

**Steps Involved in Building the Project**

1. **Data Preparation with SQL**:
   * Loaded the dataset into SQLite (in-memory).
   * Aggregated metrics: average tenure, total monthly charges, internet usage.
   * Saved processed data to processed\_data.csv.
2. **Exploratory Data Analysis (EDA)**:
   * Visualized churn distribution: 27% churn rate (1,869 customers).
   * Analyzed correlations: tenure negatively correlates with churn (-0.35); monthly charges positively correlate (0.19).
   * Generated plots: churn distribution, correlation matrix.
3. **Model Building**:
   * Built a Random Forest Classifier pipeline with preprocessing (scaling, encoding).
   * Split data: 70% train, 30% test (stratified).
   * Evaluated model: accuracy ~80%, AUC-ROC 0.82, recall for churn 0.47.
4. **Model Explainability**:
   * Used ELI5 to identify key features: one-year contracts (weight: 0.0039), online security (0.0029).
   * Used SHAP to analyze prediction impacts: fiber optic service, short tenure, and high monthly charges increase churn risk.
   * Generated SHAP summary plot and ELI5 feature importance table.
5. **Customer Segmentation**:
   * Applied K-Means clustering on tenure, monthly charges, and total charges.
   * Identified segments: At Risk (47% churn), Loyal (15% churn), Dormant (12% churn).
   * Saved segmented data to segmented\_customers.csv.
6. **Deliverables**:
   * Python notebook (CustChurn.py) with all steps.
   * PowerPoint report (Customer Churn Prediction for Telecom Industry.pptx) summarizing findings.
   * Recommendations in ppt presentations (slide -10) promote long-term contracts, enhance tech support, target at-risk customers.

**Conclusion**

The project successfully predicted customer churn with a reliable model (AUC-ROC 0.85) and identified actionable insights. Short tenure, high charges, and fiber optic service are key churn drivers, while one-year contracts and tech support reduce churn. Segmentation into At Risk, Loyal, and Dormant groups enables targeted retention strategies. Recommendations include contract incentives and service improvements. Future steps involve deploying the model in a CRM system, A/B testing retention strategies, and retraining with new data to sustain accuracy.