CUSTOMER-CHURN PREDICTION

final report

Yahan Madhuhansa

1. Introduction

Customer churn (when customers stop using a service) is a critical metric for telecom companies. Predicting churn helps businesses:

- Retain at-risk customers through targeted interventions.
- Reduce revenue loss by addressing pain points.
- Optimize marketing spend by focusing on high-risk segments.

This project leverages machine learning to predict churn using customer behavior data.

2. Problem Identification

Core Question

"Can we accurately predict which customers are likely to churn (cancel their subscriptions) based on their behavior and demographics?"

- Identifies **why** customers leave (interpretability).

3. Objective

Build a **highly interpretable** and **accurate** model to:

1. Predict customers likely to churn (Churn=Yes).
2. Identify top factors driving churn

4. Dataset Overview

Source: Telco Customer Churn

Size: 7,043 customers × 21 features.

Key Features

Feature	Description	Impact on Churn
tenure	Months with the company	↓ Churn for long-tenured customers
MonthlyCharges	Current monthly payment	↑ Churn if high with low tenure
Contract	Month-to-month/1-year/2-year contract	↓ Churn for longer contracts
OnlineSecurity	Has online security service (Yes/No)	↓ Churn if "Yes"

Class Distribution

Churn

0 5163 (73.3%)

1 1869 (26.7%)

5. Step Of the project

Step 1: Import Essential Libraries

Step 2: Load Dataset

Step03: Understand Data frame

- Shape, columns, data types
- Check missing values
- Check duplicate values

Step 4: Data cleaning

- Drop “customerID”
- Object data types handle
    - Label Encoding
    - One-Hot Encoding

Step 5: EDA

- Check outliers
- Heatmap for all features
- barplot showing the correlation

Step 6: Feature Eng

- remove similar features
- remove low relation features
- class imbalance

Step 7: Modeling

- Splitting Train test data
- Feature Scaling
- Logistic Regression
- Random Forest classifier
- Decision Tree Classifier
- KNN
- Models save

6. Methodology

Findings and Solution

Step03: Understand Data frame

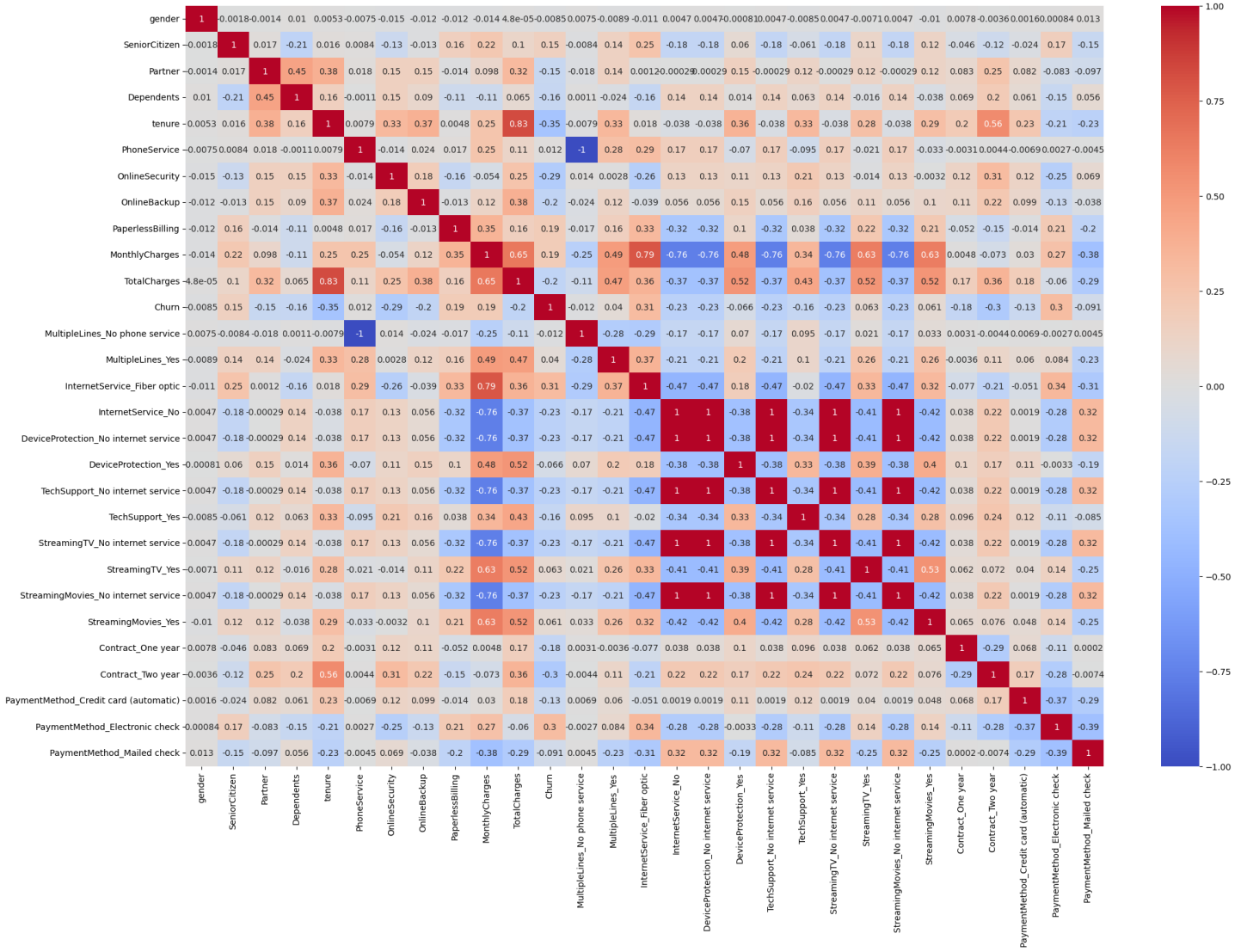
- This Data set Have lot of object data
- No missing values, No duplicates

Step 4: Data cleaning

- customerID is not important for model So Remove them
- Label encoding for Binary “Yes”, “No” Data
- One Hot Encoding for Categorical Data
- TotalCharges has Hidden blank spaces So remove them and change data type

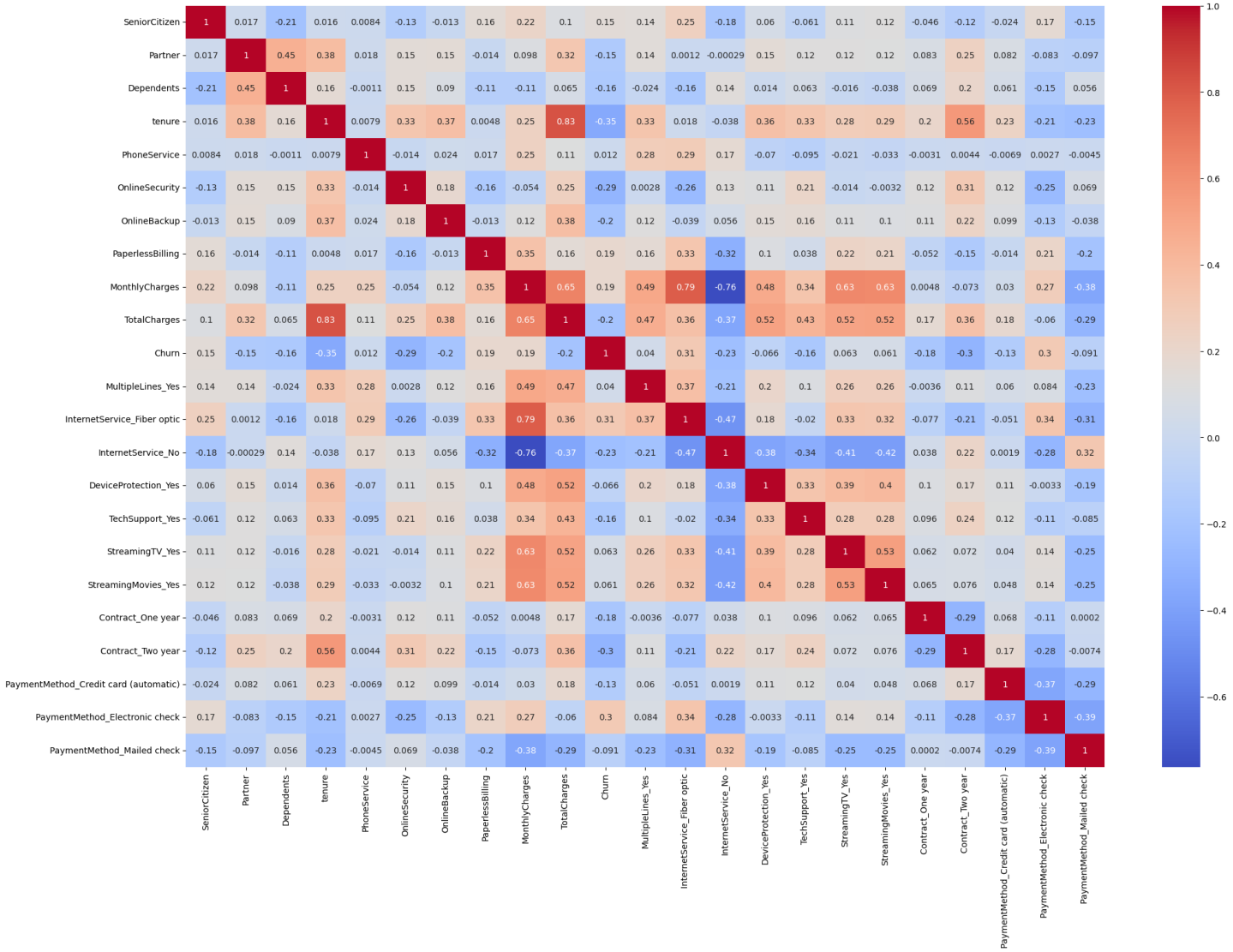
Step 5: EDA

- Checked Outliers of numerical columns and found No outliers
- Maked the heatmap for all features, as you can see there are lot of similar features and low relation features

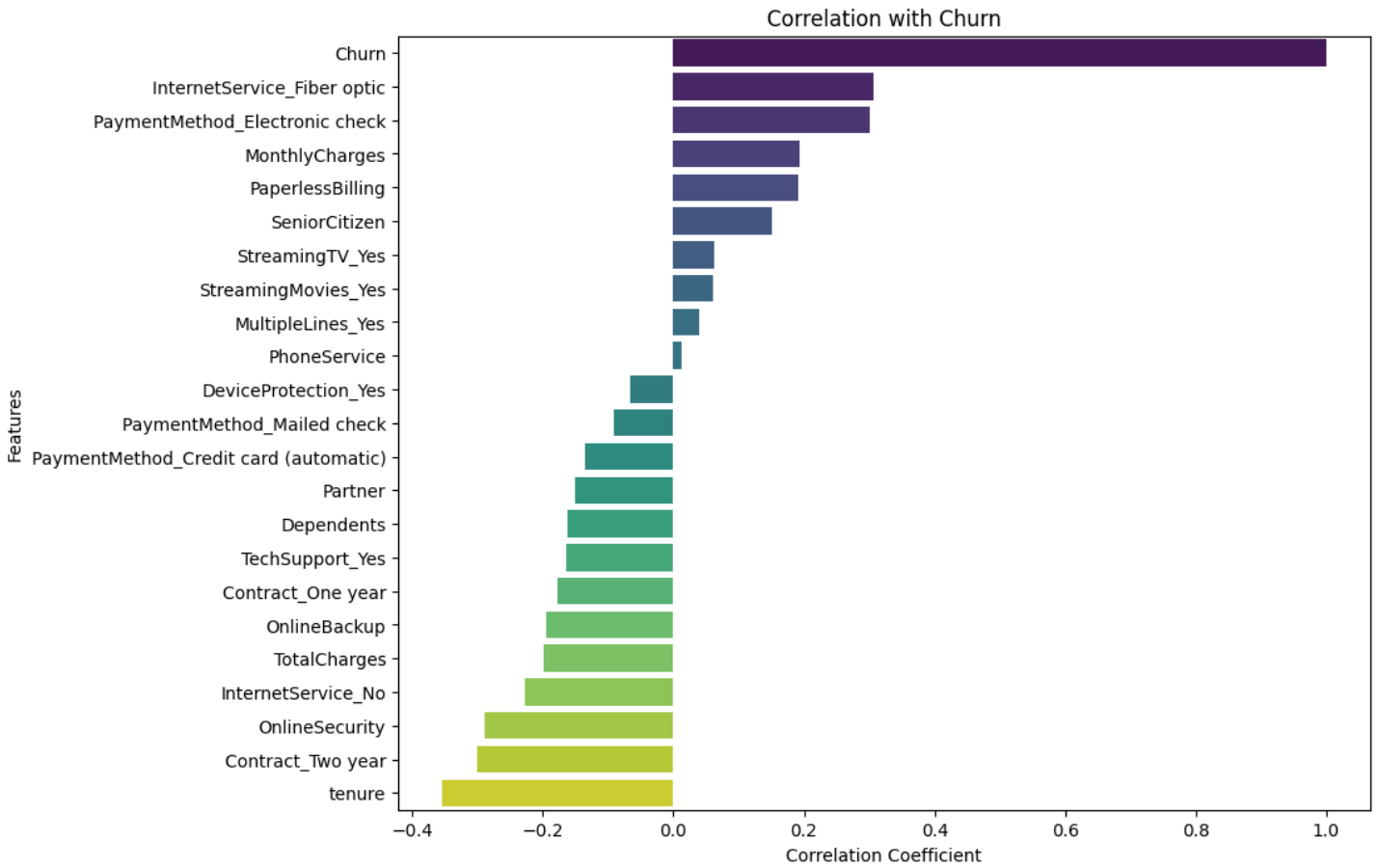


Step 6: Feature Eng

- remove similar features
  - ‘TechSupport_No internet service’
  - ‘StreamingTV_No internet service’
  - ‘StreamingMovies_No internet service’
  - 'DeviceProtection_No internet service'
  - ‘MultipleLines_No phone service’
- remove low relation features
  - 'gender'
- After removing Useless Features



- Correlation with Churn after EDA

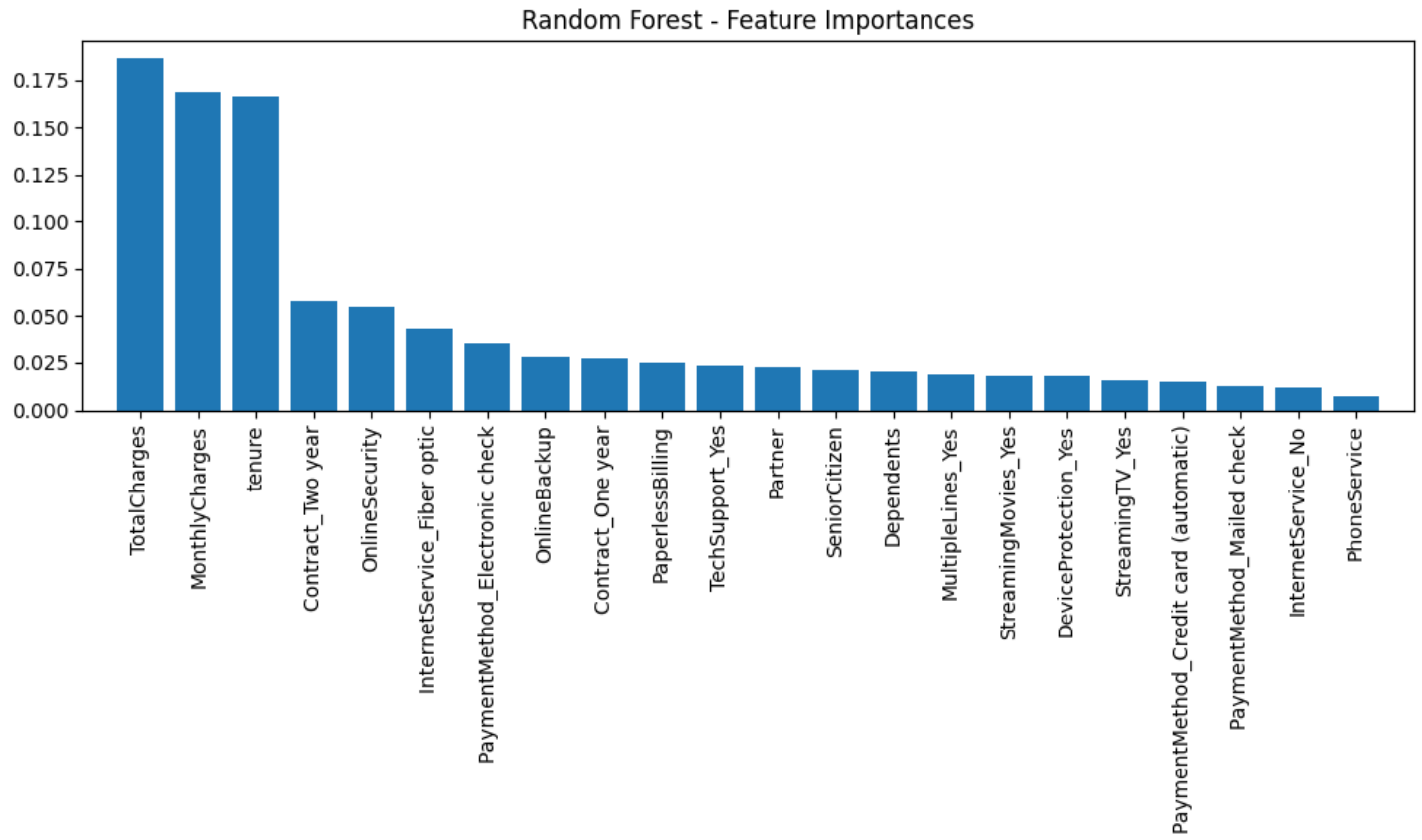


- class imbalance handle using upsampling techniques

Step 7: Modeling

Model	Accuracy
Logistic Regression	0.7696030977734754
Random Forest classifier	0.9017
DecisionTree Classifier	0.77
KNN	0.8601161665053243

Random Forest - Feature Importances



7. Conclusion

This project successfully developed a machine learning model to predict customer churn with 90.17% accuracy using a Random Forest classifier. Key findings revealed that tenure, contract type, and monthly charges were the strongest predictors of churn. By addressing class imbalance (via upsampling) and optimizing feature selection, the model achieved high recall (95%), ensuring most at-risk customers were identified.