

Vodafone Customer Churn Prediction

Hrishinandan N
BCB 304

Problem Statement

Why is churn a problem for the business?

Goal

To build a predictive model that accurately flags customers at high risk of churning.

Data Overview & Cleaning

Data Source: Kaggle ([click here](#))

Initial State: 7,043 records, 23 features

Key Cleaning Steps:

- Handling Missing values (median/mode)
- Converting TotalCharges into numeric
- Removing non-predictive columns (customerID/Location)

Feature Engineering

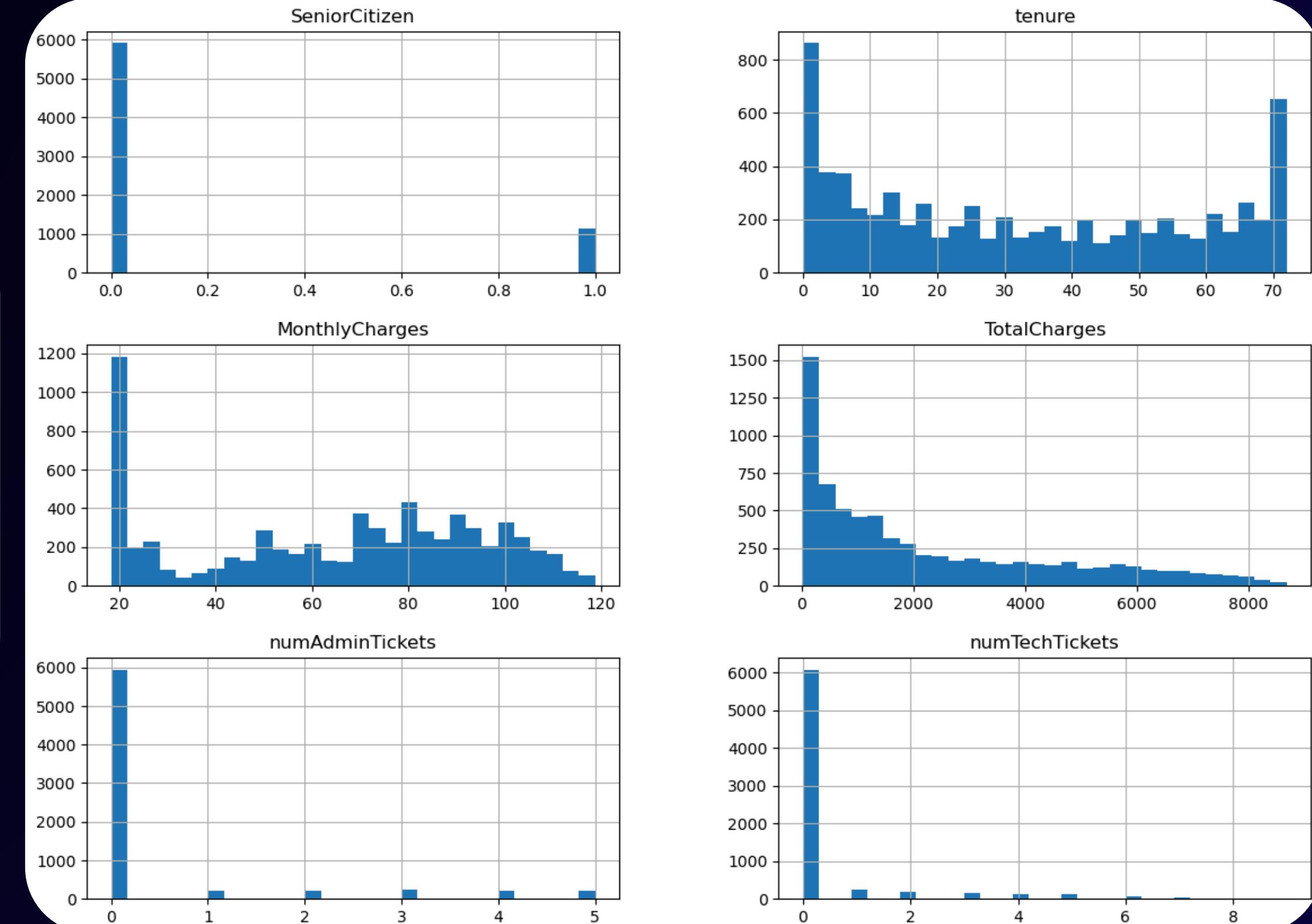
Categorical Handling

One-Hot Encoding on Contract, PaymentMethod
and InternetService.

Label Encoding on all other binary features

Exploratory Data Analysis

Numerical Features



Numerical features vs Churn

Box Plot Insights

Strong Predictors

- Low tenure
- Low TotalCharges
- High MonthlyCharges
- SeniorCitizen status
- Low numTickets

Weak Predictors

numAdminTicket

Categorical feature vs Churn

Count Plot Insights

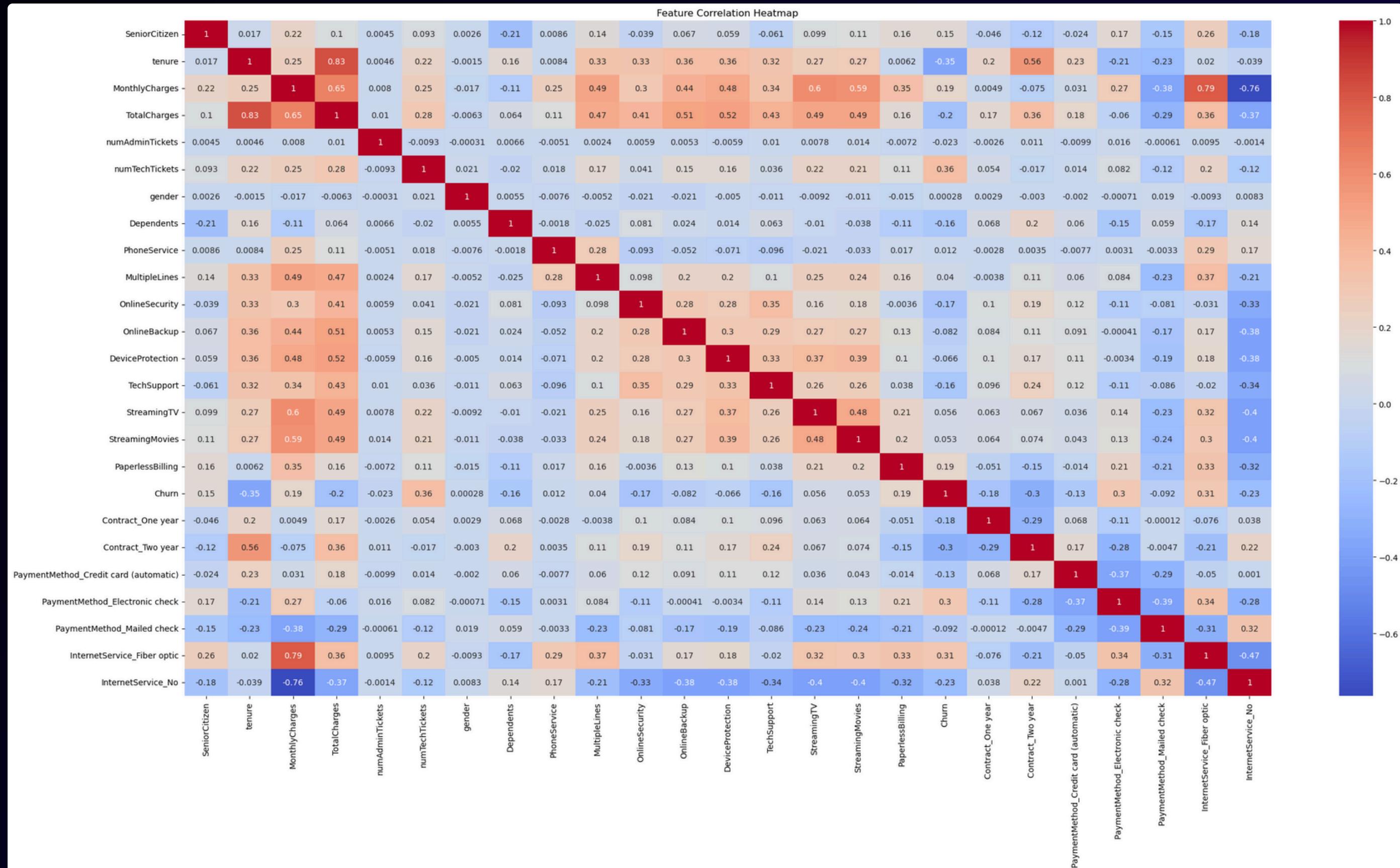
Strong Predictors

- Short term contracts
- E - Check payment
- Fiber Optic Internet
- No Dependents
- No TechSupport
- No OnlineServices
- Paperless Billing

Weak Predictors

- Gender
- PhoneService
- StreamingTV
- StreamingMovies
- MultipleLines

Correlation Analysis



Feature Selection

14 selected

- Senior Citizen
- Tenure
- Monthly Charges
- Total Charges
- Num TechTickets
- Dependents
- Internet Service
- Online Security
- Online Backup
- Device Protection
- Tech Support
- Paperless Billing
- Contract
- Payment Method

Model Selection & Metrics

Algorithms Tested

- Logistic Regression
- K Nearest Neighbors
- Decision Tree Classifier
- Support Vector Machine

Metrics

- | | |
|-------------|------------|
| • Accuracy | • Recall |
| • Precision | • F1 Score |

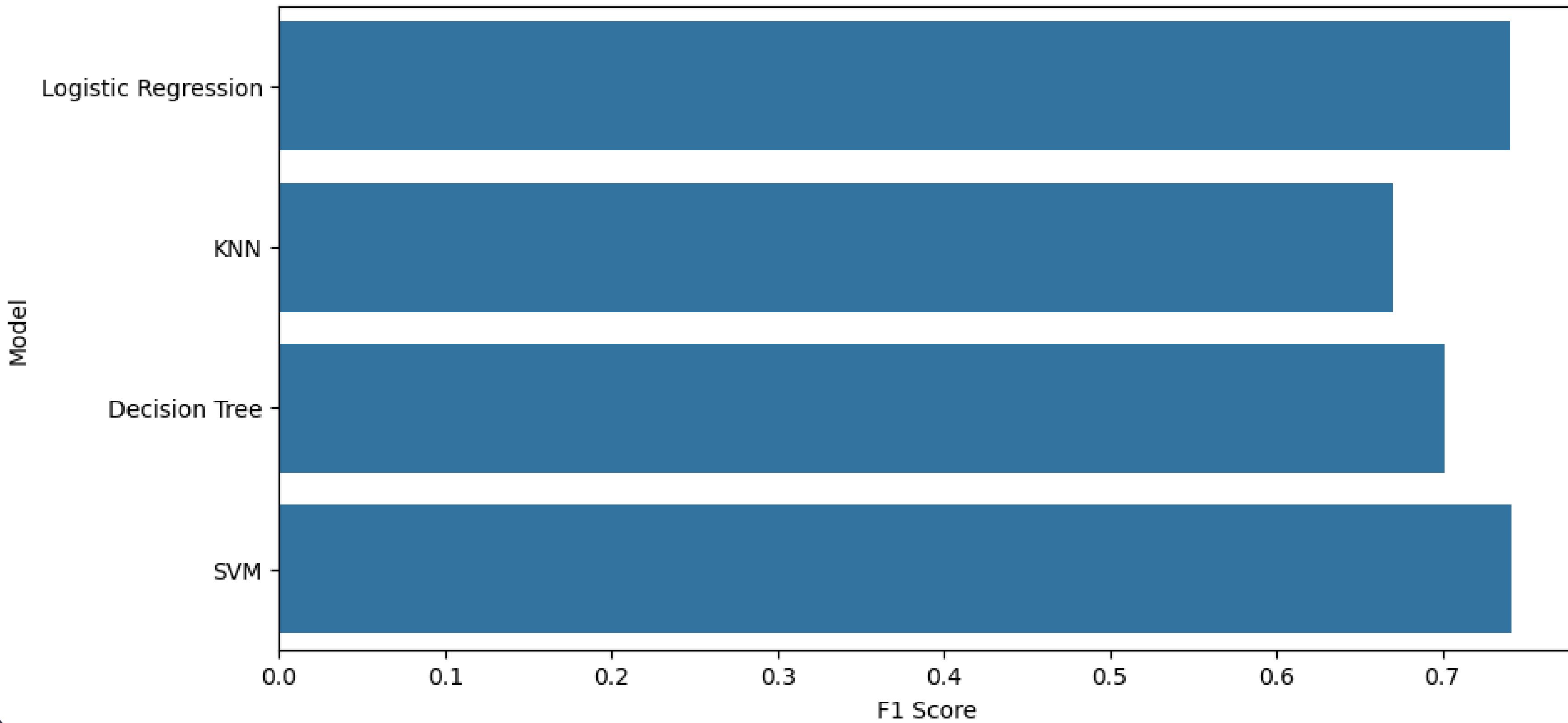
F1 score balances Precision and Recall, making it the primary metric.

Model Performance Comparison

	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.865153	0.753463	0.729223	0.741144
KNN	0.832505	0.699708	0.643432	0.670391
Decision Tree	0.839603	0.690909	0.713137	0.701847
SVM	0.865862	0.755556	0.729223	0.742156

The SVM model offers the best trade-off, correctly identifying high-risk customers about **74%** of the time while maintaining a balanced rate of false positives and false negatives.

Model Comparison



Thank You