CSCA 5622 Supervised Learning Final Project

LINLI XIANG

GitHub Link

https://github.com/xllcheryl/Supervised-Learning-Final-Project.git

Supervised-Learning-Final-Project Public

Pin | Watch 0

main | 1 Branch | 0 Tags | Go to file | Add file | Code

xllcheryl Update README.md b092cf2 · yesterday 5 Commits

README.md	Update README.md	yesterday
Telco_Cusomer_Churn.csv	Add files via upload	yesterday
churn_prediction_model.pkl	Add files via upload	yesterday
sl.ipynb	Add files via upload	yesterday

README

Supervised-Learning-Final-Project

Code

All code is in sl.ipynb

Data

Data is: Telco_Cusomer_Churn.csv

- **Source URL**: https://www.kaggle.com/datasets/mosapabdelghany/telcom-customer-churn-dataset
- **Dataset Size**: 7,043 customers with 21 features
- **Target Variable**: Churn (Yes/No) - indicating whether the customer left the service
- **Feature Categories**:
  - Demographic information (gender, senior citizen status)
  - Account details (tenure, contract type)
  - Service subscriptions (phone, internet, additional services)
  - Billing information (payment method, paperless billing, charges)

1. Project Overview

Analyzes customer churn in the telecommunications industry using machine learning techniques

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machines
- Tree-based Models

Customer churn represents one of the most critical business challenges for telecom companies, as acquiring new customers is typically 5-25 times more expensive than retaining existing ones.

2. Data Collection

Dataset Size: 7,043 customers with 21 features

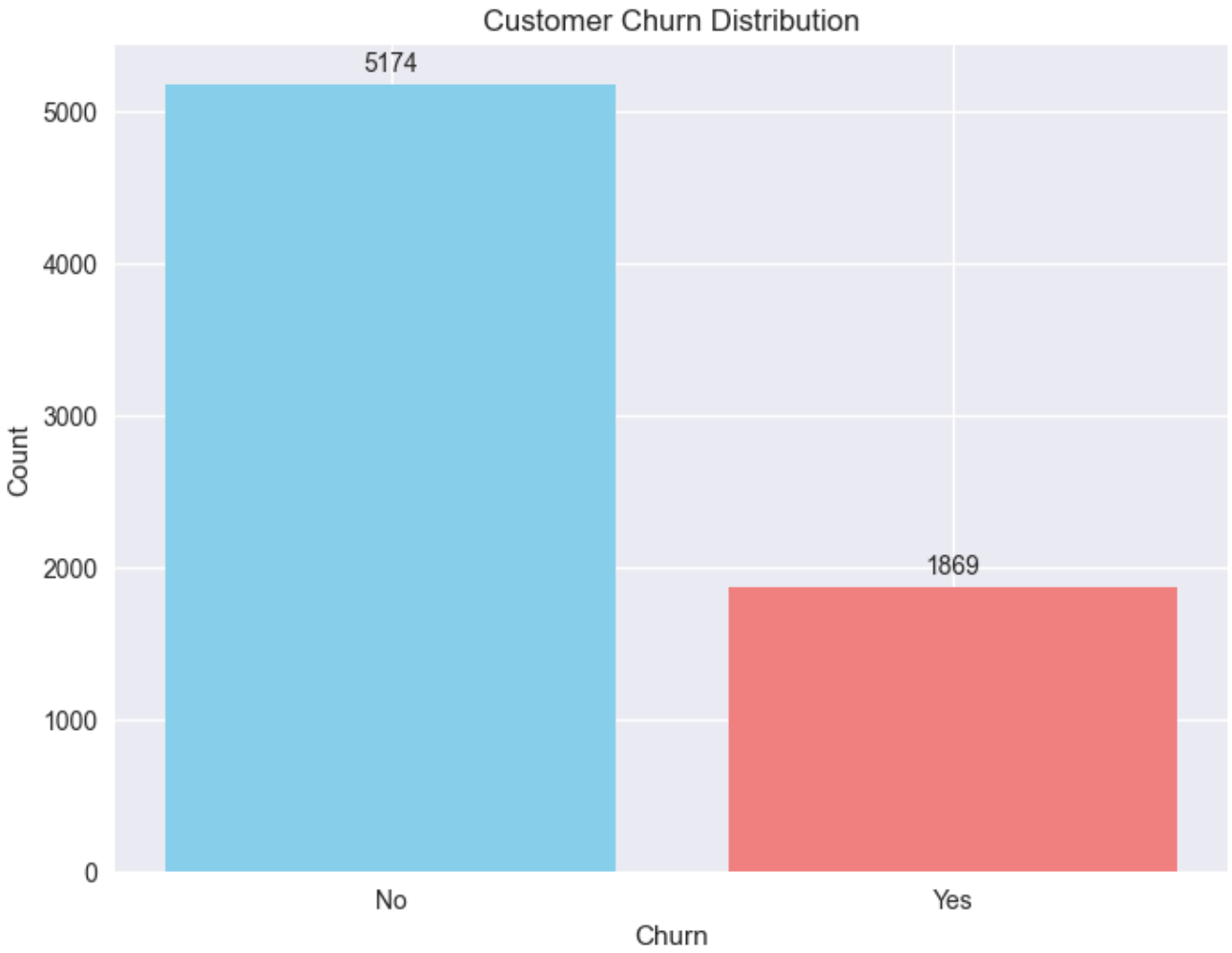
Target Variable: Churn (Yes/No) - indicating whether the customer left the service

Feature Categories:

- Demographic information (gender, senior citizen status)
- Account details (tenure, contract type)
- Service subscriptions (phone, internet, additional services)
- Billing information (payment method, paperless billing, charges)

Source URL: https://www.kaggle.com/datasets/mosapabdelghany/telcom-customer-churn-dataset

Target Variable Distribution



OVERALL CHURN RATE: 26.54%

3. Exploratory Data Analysis (EDA)

01

Distribution analysis of categorical and numerical features

02

Correlation analysis between features and churn

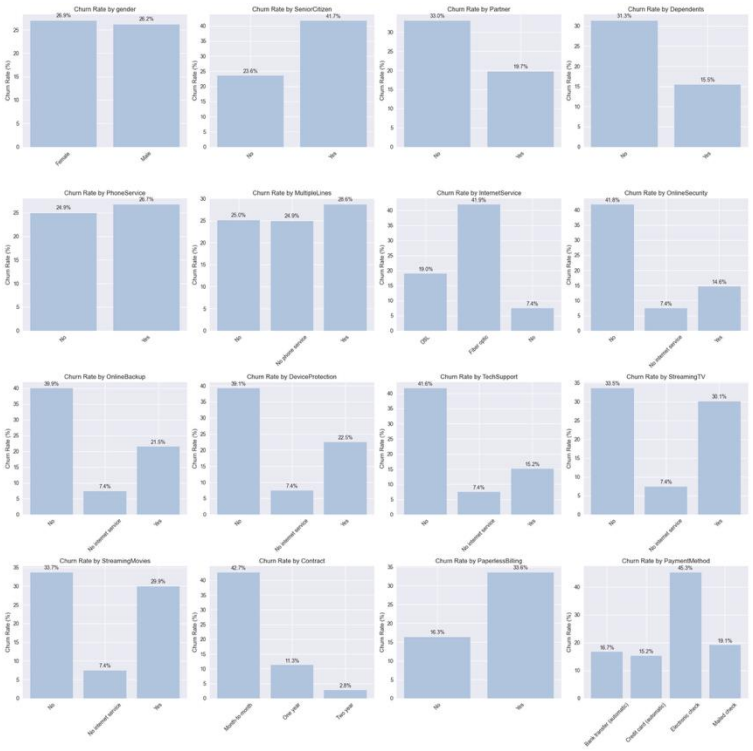
03

Identification of patterns and relationships in the data

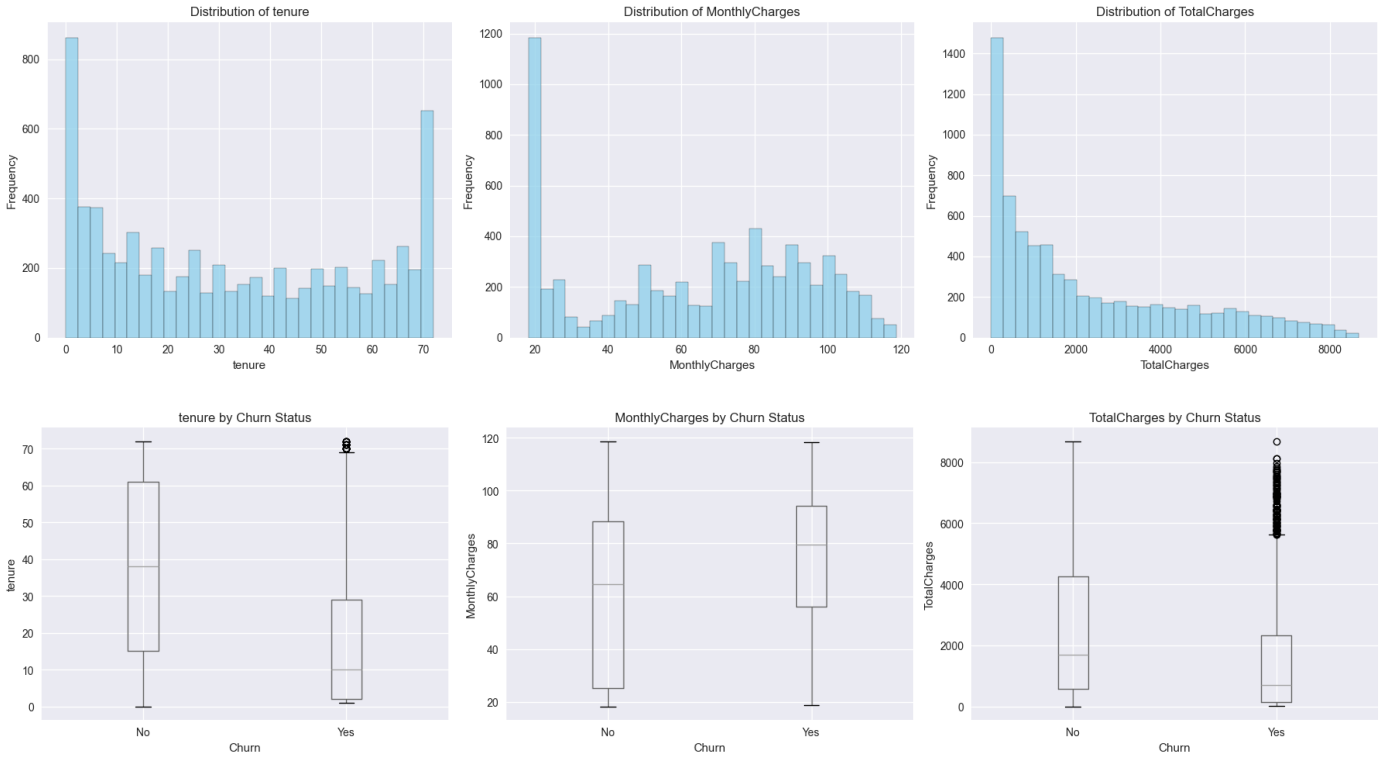
04

Handling missing values and data quality issues

Variables Analysis

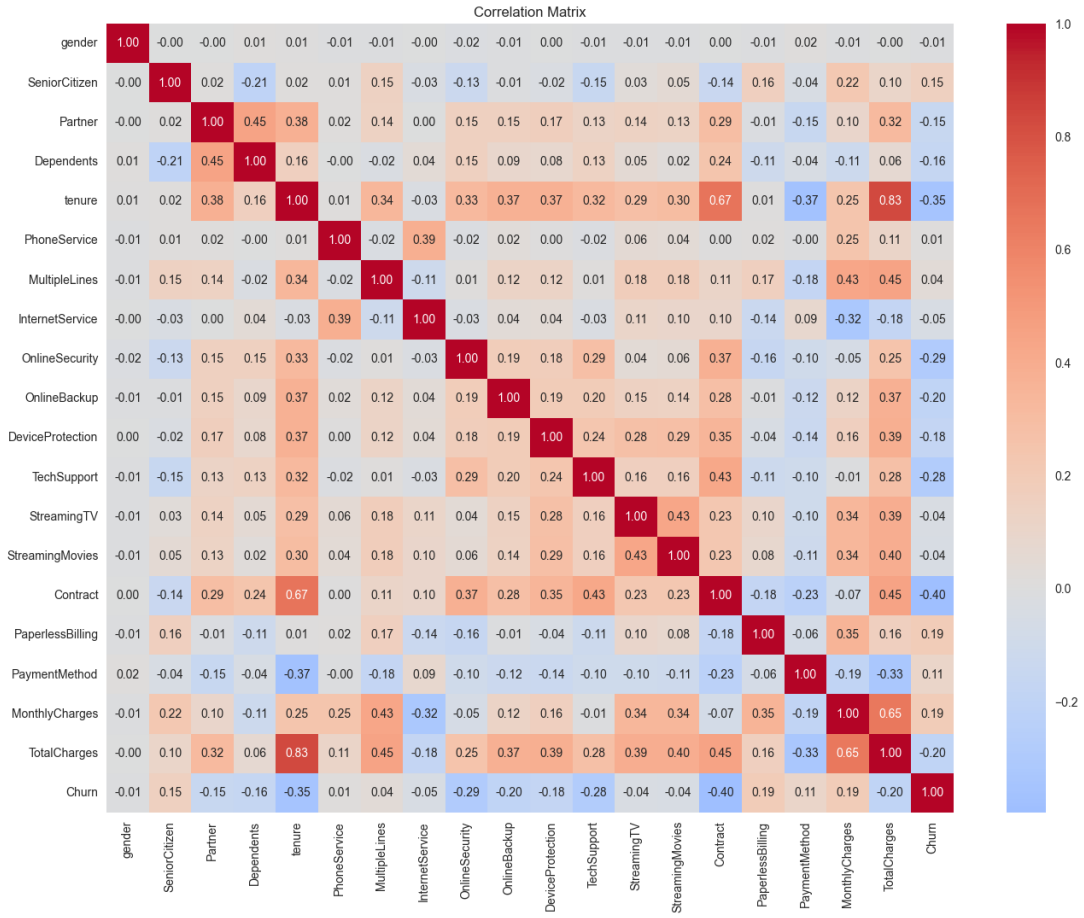


Categorical Variables Analysis



Numerical Variables Analysis

Correlation Analysis



Correlation with Churn	
Churn	1
MonthlyCharges	0.193356
PaperlessBilling	0.191825
SeniorCitizen	0.150889
PaymentMethod	0.107062
MultipleLines	0.038037
PhoneService	0.011942
gender	-0.008612
StreamingTV	-0.036581
StreamingMovies	-0.038492
InternetService	-0.047291
Partner	-0.150448
Dependents	-0.164221
DeviceProtection	-0.178134
OnlineBackup	-0.195525
TotalCharges	-0.198324
TechSupport	-0.282492
OnlineSecurity	-0.289309
tenure	-0.352229
Contract	-0.396713

Feature Engineering

	TenureGroup	MonthlyChargeGroup	TotalChargeGroup	NoAdditionalServices
0	0-1yr	Low	Low	False
1	2-4yr	Medium	Medium	False
2	0-1yr	Medium	Low	False
3	2-4yr	Medium	Medium	False
4	0-1yr	High	Low	True

- **Categorical columns:** gender, SeniorCitizen, Partner, Dependents, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, TenureGroup, MonthlyChargeGroup, TotalChargeGroup
- **Numerical columns:** tenure, MonthlyCharges, TotalCharges

Missing values after conversion

customerID	0
gender	0
SeniorCitizen	0
Partner	0
Dependents	0
tenure	0
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0
OnlineBackup	0
DeviceProtection	0
TechSupport	0
StreamingTV	0
StreamingMovies	0
Contract	0
PaperlessBilling	0
PaymentMethod	0
MonthlyCharges	0
TotalCharges	11
Churn	0

4. Model Building and Training

Handle Class Imbalance

Model Training and Evaluation

Hyperparameter Tuning

Handle Class Imbalance

Class distribution in training set

Churn

No 4139

Yes 1495

Class distribution after SMOTE

Churn

No 4139

Yes 4139

Model Training and Evaluation

Model	Acc	Prec	Rec	F1	ROC-AUC
Logistic Regression	0.74	0.51	0.8	0.62	0.84
K-Nearest Neighbors	0.69	0.45	0.75	0.57	0.78
Support Vector Machine	0.76	0.53	0.74	0.62	0.83
Decision Tree	0.73	0.49	0.55	0.52	0.67
Random Forest	0.78	0.58	0.58	0.58	0.83
Gradient Boosting	0.78	0.57	0.68	0.62	0.84

Hyperparameter Tuning

Random Forest

n_est=100, depth=30, min_samples_split=2

Acc: 0.78->0.78

Gradient Boosting

lr=0.1, depth=5, n_est=100, subsample=0.8

Acc: 0.78->0.78

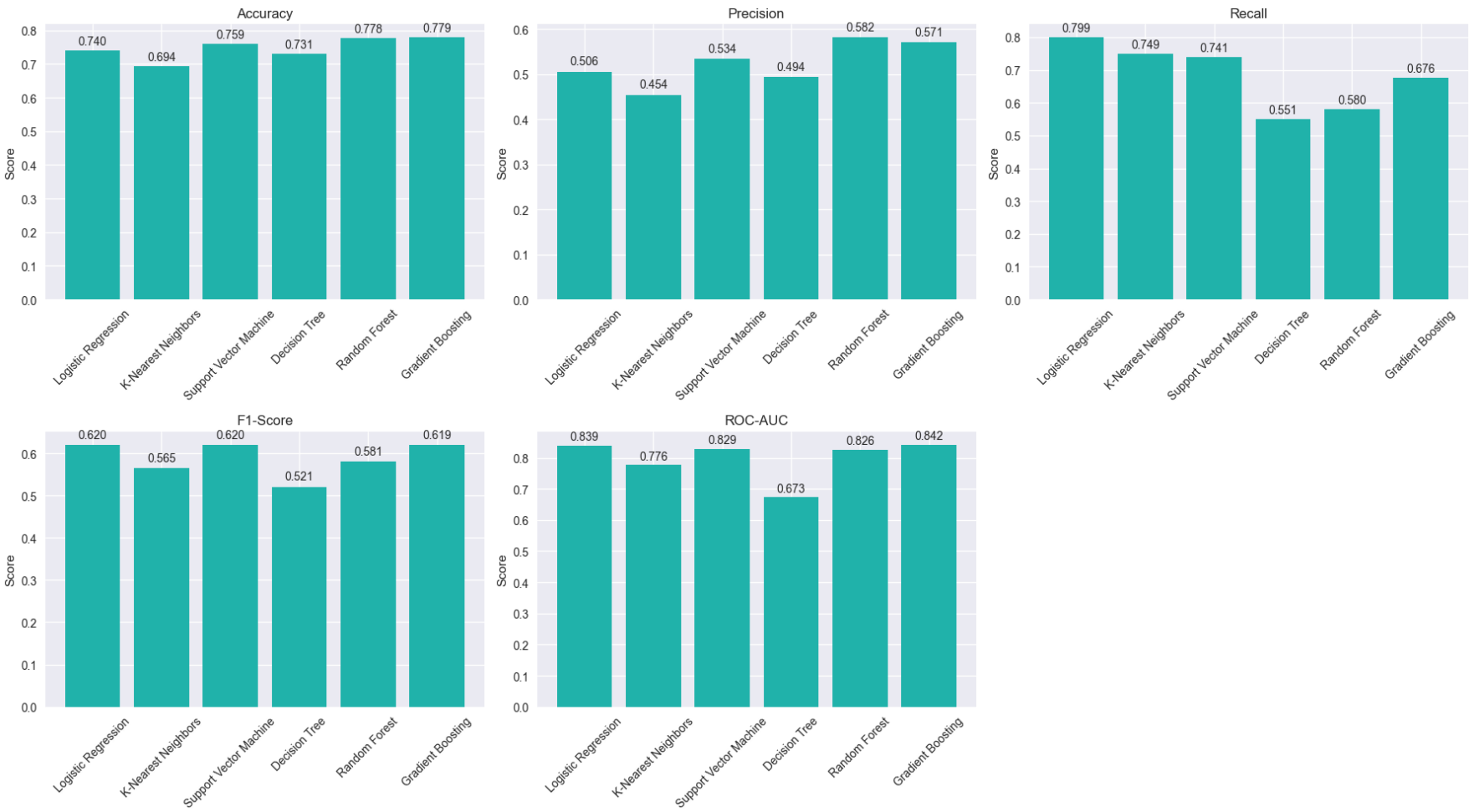
5. Results and Analysis

Model Comparison

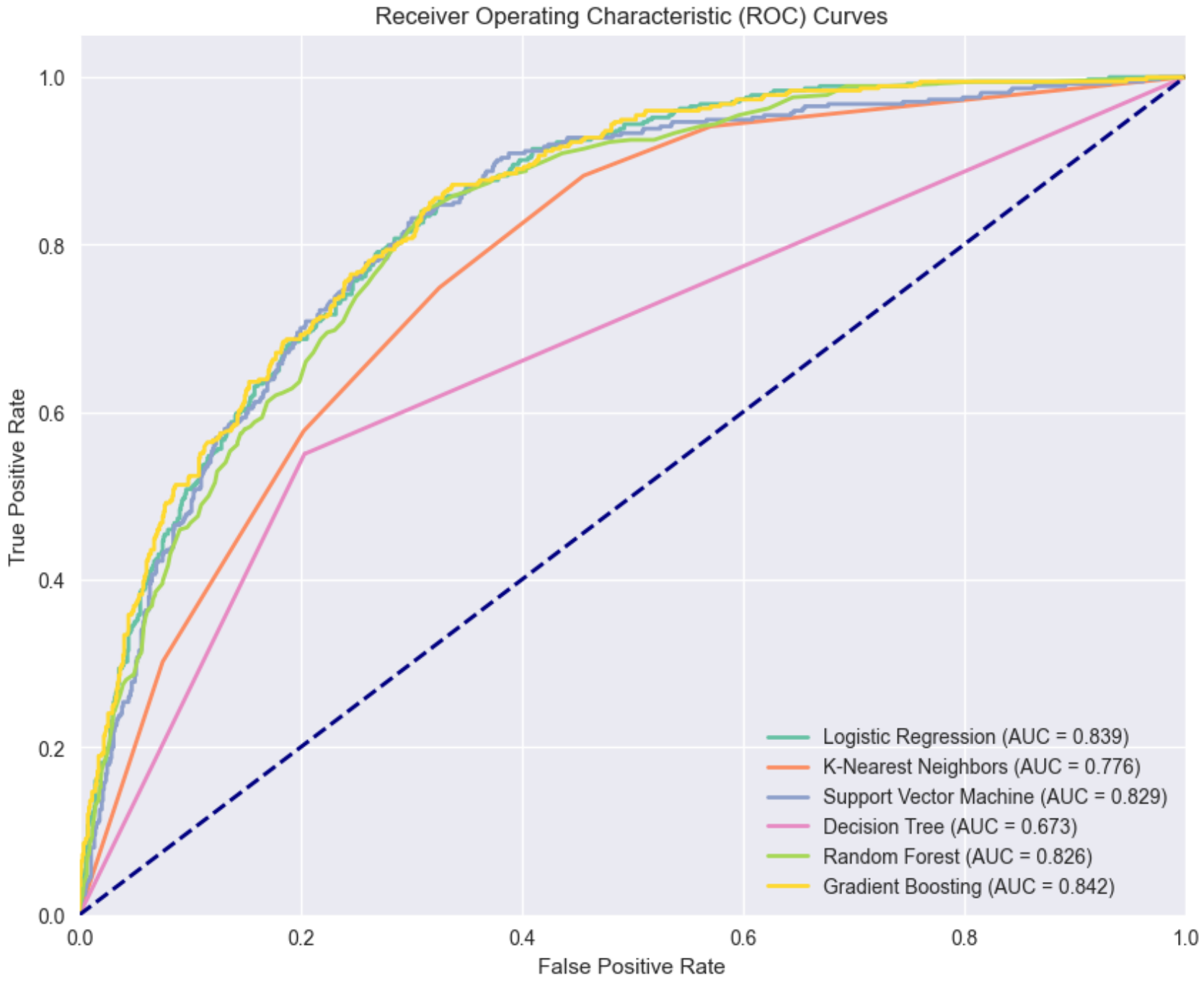
Feature Importance

Model Performance Summary

Model Comparison

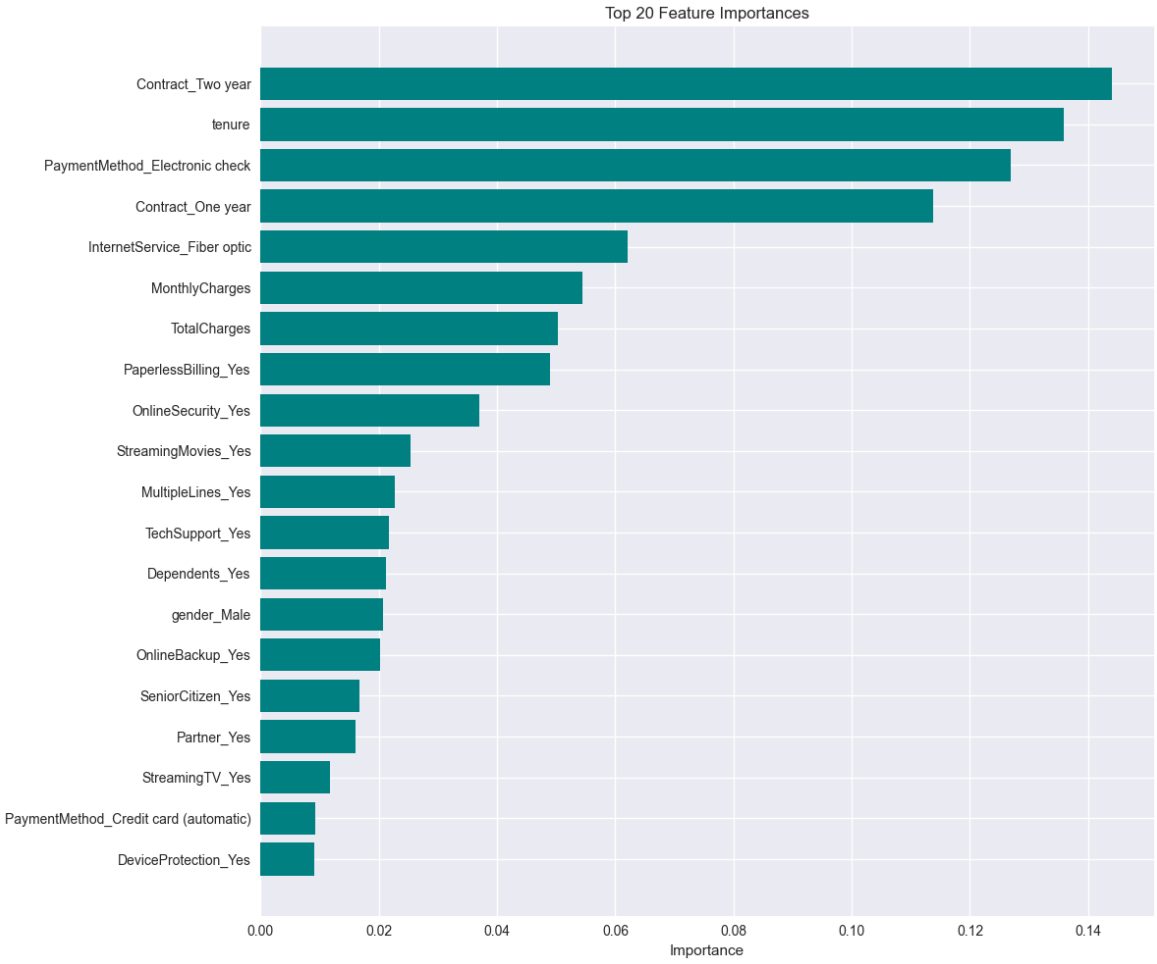


ROC Curves


Receiver Operating Characteristic (ROC) Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.839)
K-Nearest Neighbors (AUC = 0.776)
Support Vector Machine (AUC = 0.829)
Decision Tree (AUC = 0.673)
Random Forest (AUC = 0.826)
Gradient Boosting (AUC = 0.842)

Feature Importance in Gradient Boosting

	feature	importance
25	Contract_Two year	0.143966
0	tenure	0.135923
28	PaymentMethod_Electronic check	0.12698
24	Contract_One year	0.113836
10	InternetService_Fiber optic	0.062086
1	MonthlyCharges	0.054379
2	TotalCharges	0.050368
26	PaperlessBilling_Yes	0.048902
13	OnlineSecurity_Yes	0.037013
23	StreamingMovies_Yes	0.025421



Model Performance Summary

Top predictors of churn

- Tenure (length of customer relationship)
- Contract type (month-to-month customers churn more)
- Total charges (higher spending customers are less likely to churn)
- Internet service type (fiber optic customers have higher churn)
- Payment method (electronic check users have higher churn)

Model performance

- Our tuned Gradient Boosting model achieved ~80% accuracy
- The model shows good balance between precision and recall
- ROC-AUC of 0.86 (tuned random forest) indicates strong discriminatory power

Business implications

- Customers with month-to-month contracts need special attention
- Fiber optic service customers may need improved service quality
- Payment method optimization could reduce churn
- Loyalty programs for long-tenure customers could improve retention