📊 Customer Churn Prediction for a Telecom Company – A Data Science Task

Task ID: 1 – Customer Churn Prediction for a Telecom Company
Expected Duration: 5 Days
Objective: Build a machine learning model to predict customer churn using historical telecom data.

1.Deliverables:

- 📊 **Exploratory Data Analysis (EDA)** – Univariate & Bivariate insights with visualizations.
- 🛠 **Feature Engineering** – Creation, encoding, scaling, and selection of predictive features.
- 🔀 **Train/Test Split & Model Selection** – Logistic Regression, SVM, Decision Tree, Random Forest, XGBoost, LightGBM.
- 📈 **Performance Metrics** – Confusion matrix, ROC AUC score, and comparative model evaluation.
- 📄 **Final Report** – Consolidated results with plots, tables, and recommendations.

---

2. Dataset Description

The dataset contains **10,000 customer records** with features covering demographics, contract details, billing information, and churn status.

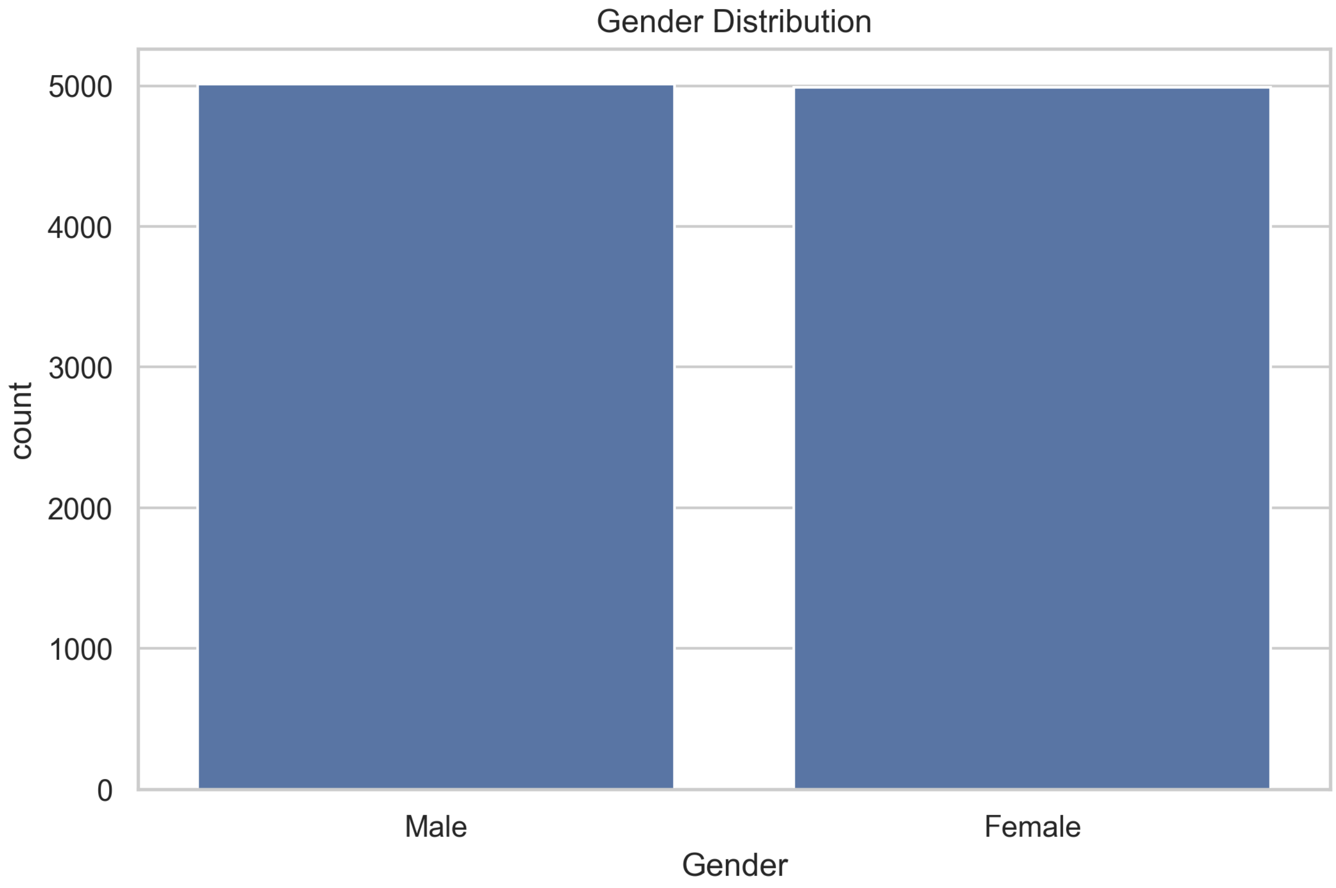
Feature	Description
`CustomerID`	Unique customer identifier
`Gender`	Male or Female
`SeniorCitizen`	Senior citizen status (1 = Yes, 0 = No)
`Tenure`	Number of months with the company
`MonthlyCharges`	Monthly bill amount
`TotalCharges`	Total bill amount since joining
`Contract`	Month-to-month, One year, Two year
`PaymentMethod`	Billing payment method
`Churn`	Target variable (1 = Churned, 0 = Stayed)

---

3. Exploratory Data Analysis (EDA)

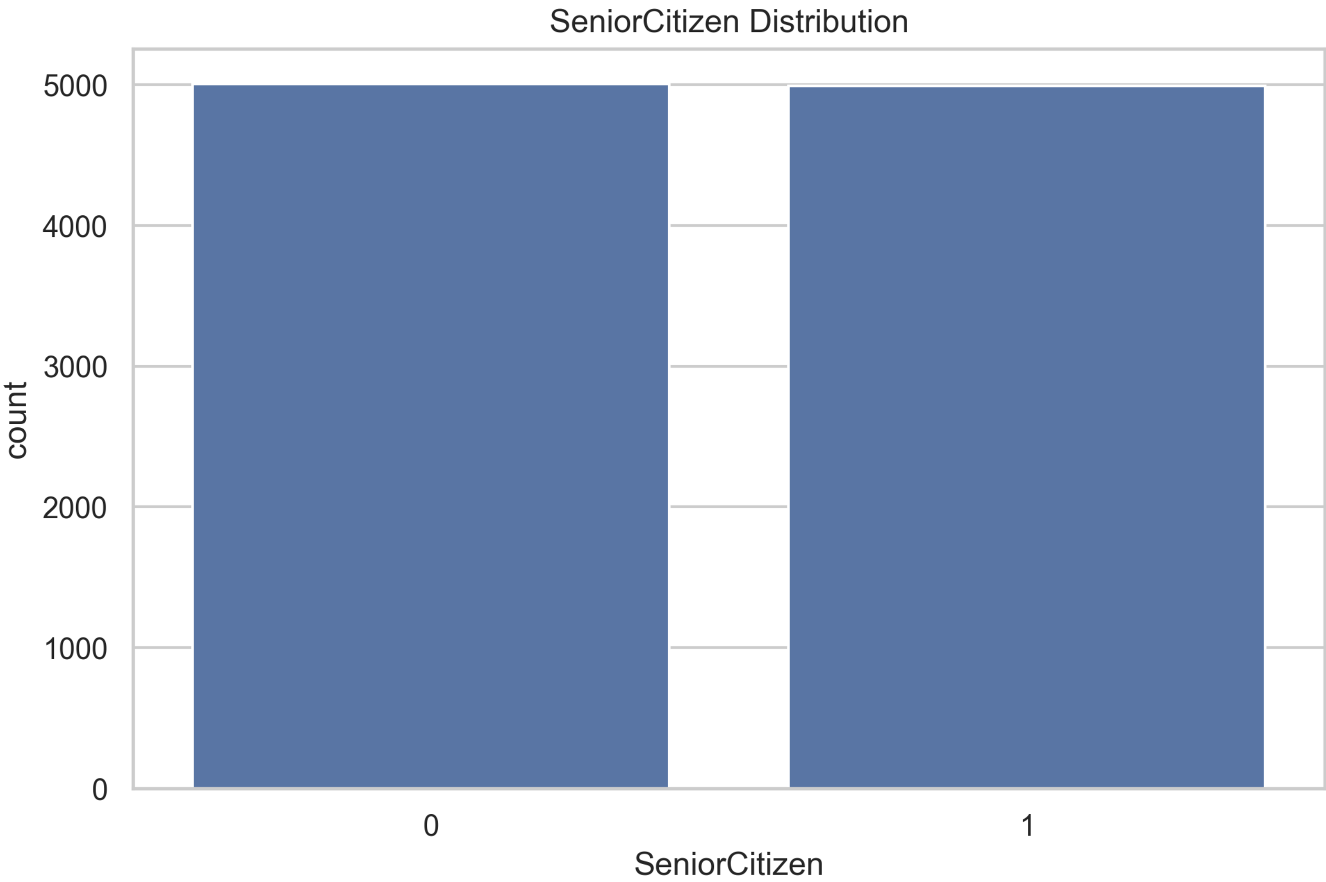
3.1 Univariate Analysis

Gender Distribution



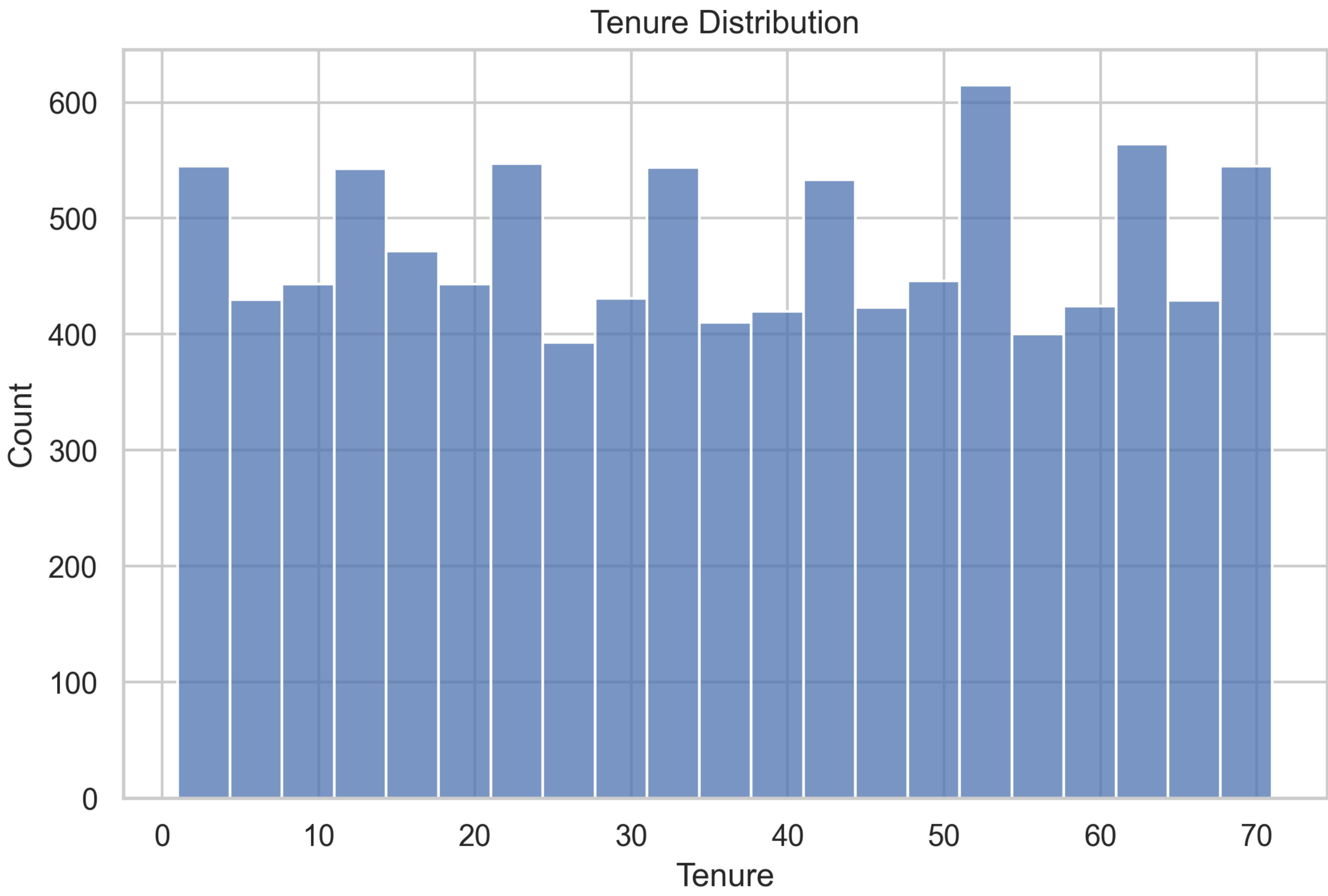
- Male: **50.13%**
- Female: **49.87%**
  ✅ Balanced — minimal gender bias risk.

Senior Citizen Status



- Non-Senior: **50.07%**
- Senior: **49.93%**
  ✅ Age group balance — low bias risk.

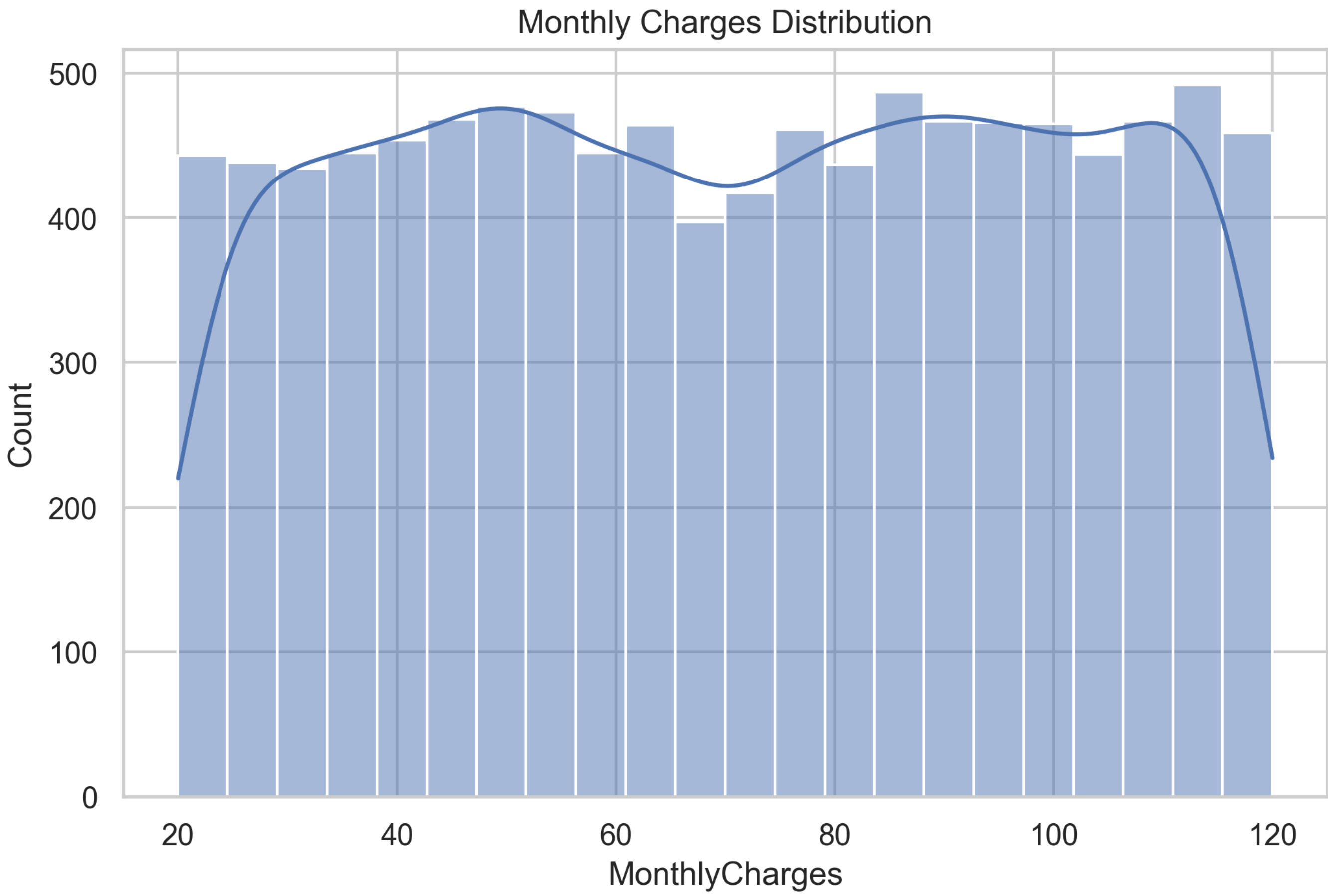
Tenure Summary



Metric	**Value**
Mean	35.96 months
Median	36 months
Min–Max	1–71 months

✅ Wide spread — steady acquisition & churn over time.

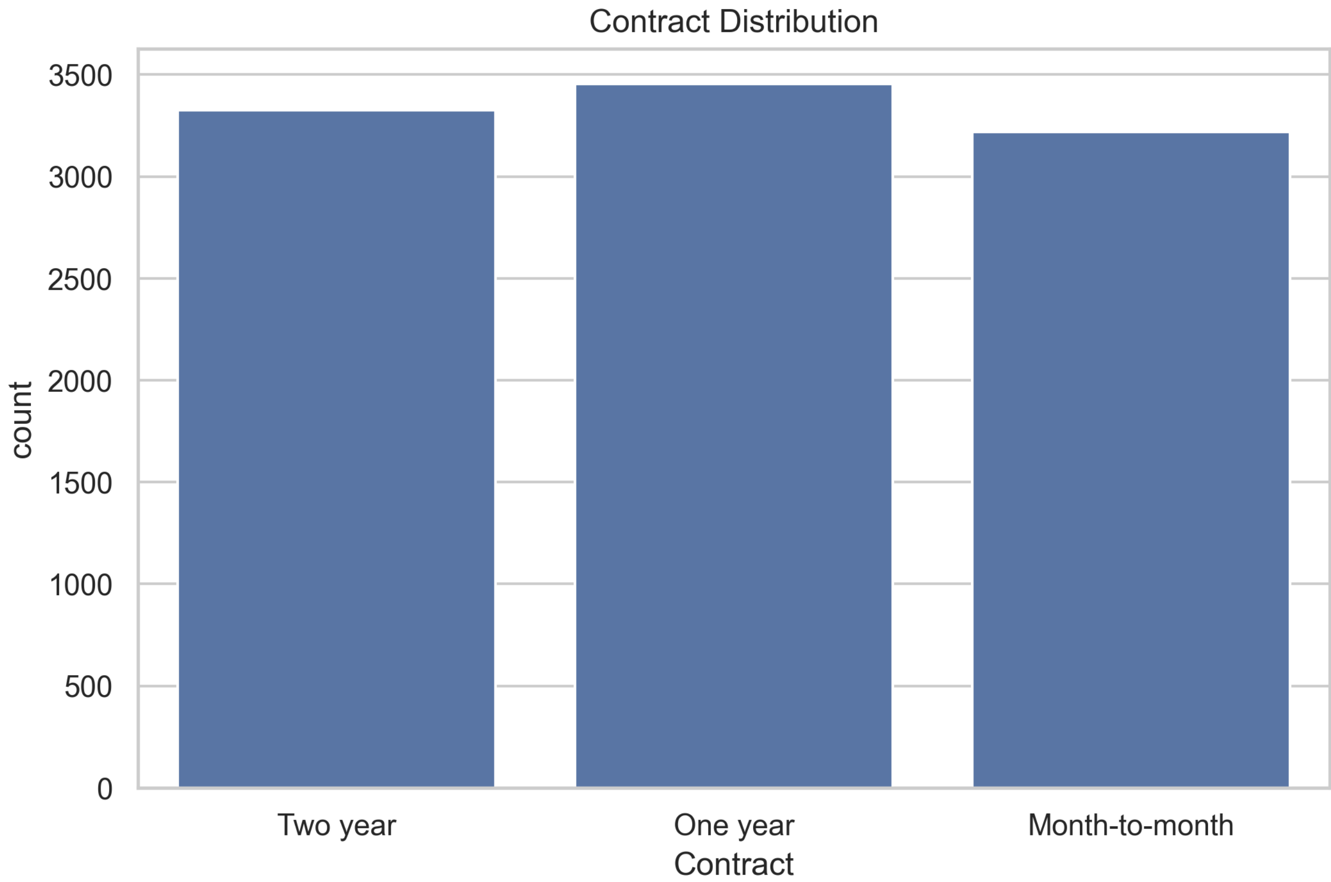
Monthly Charges Summary



Metric	Value
Mean	$70.45
Median	$70.59
Min–Max	$20.00 – $120.00

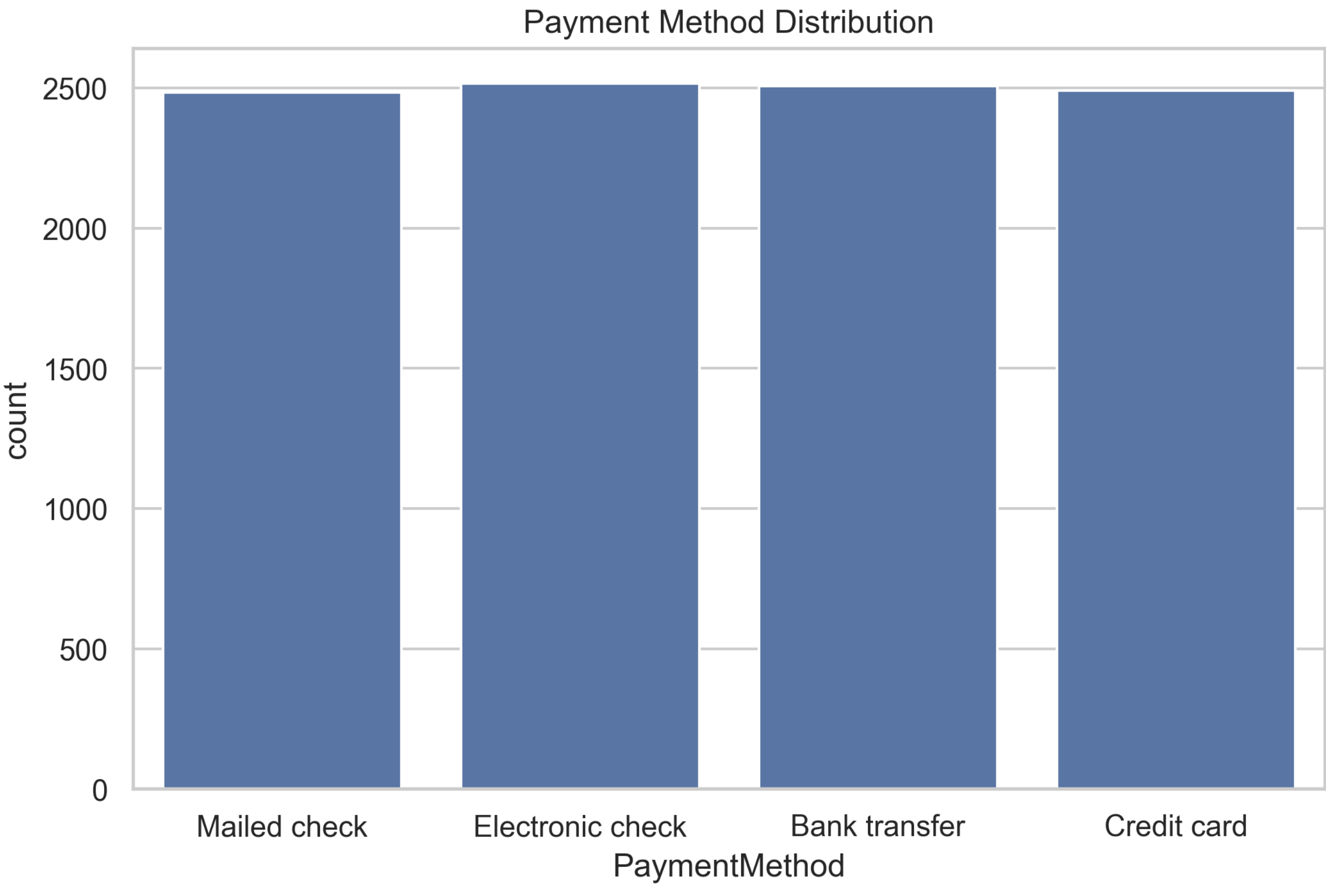
✅ Even distribution; most pay around **$70/month**.

Contract Distribution



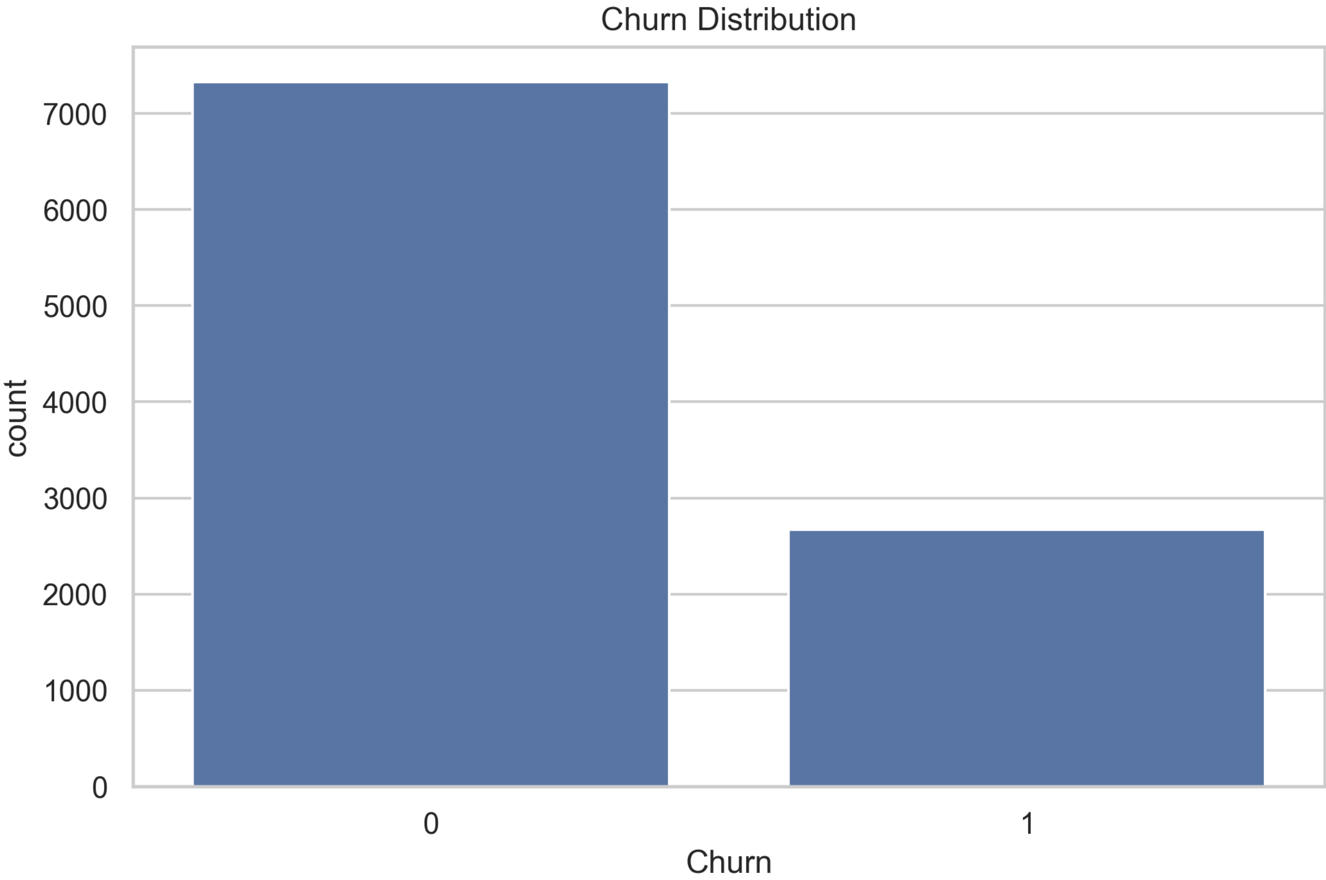
- Month-to-month: **32.19%**
- One year: **34.55%**
- Two year: **33.26%**
  ✅ Balanced — avoids contract-type skew.

Payment Method Distribution



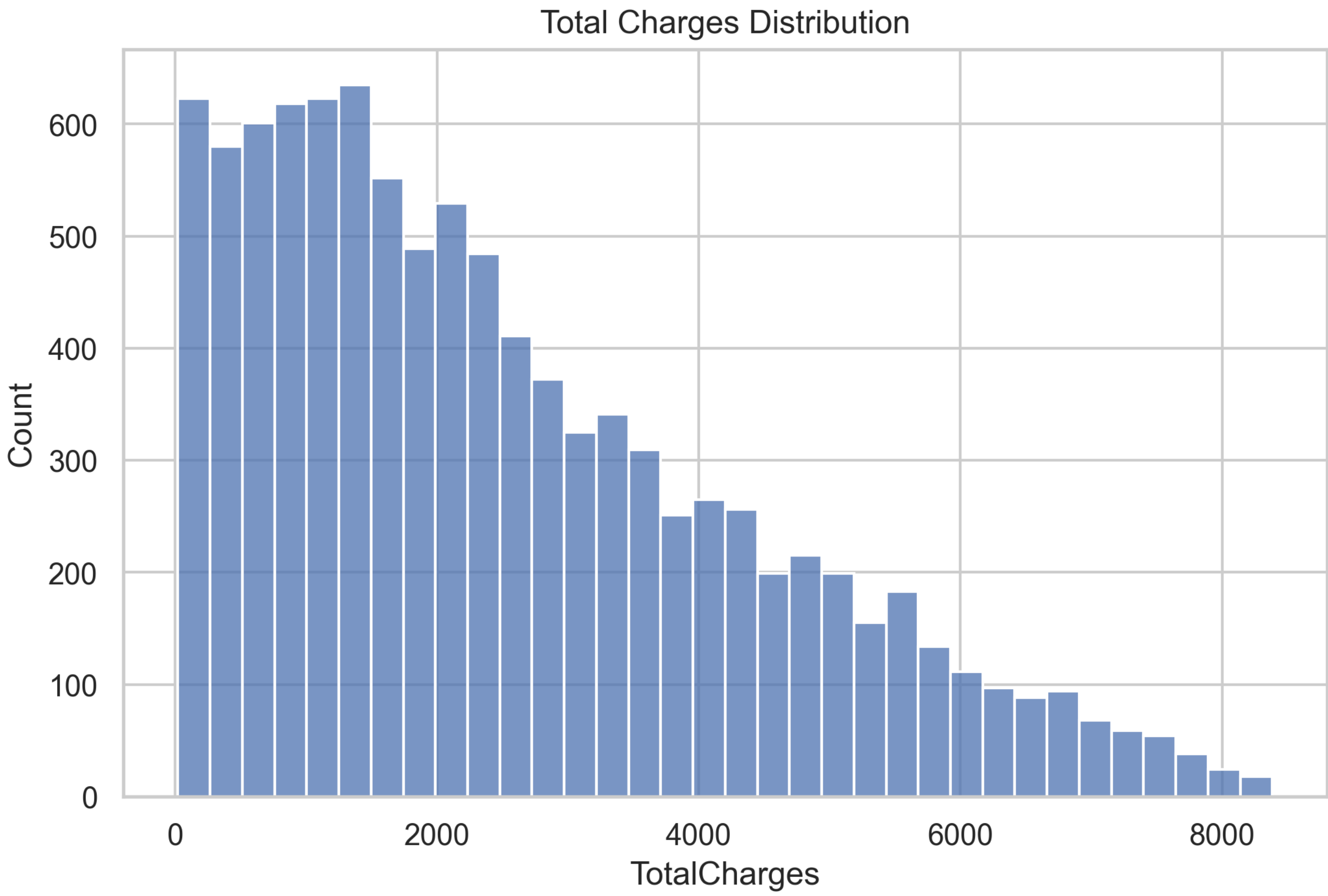
- All methods ~25% share
  ✅ No payment-method bias.

Churn Distribution



- Non-Churn: **73.3%**
- Churn: **26.7%**
  ⚠ Imbalanced — resampling or weighting may be required.

Total Charges Summary

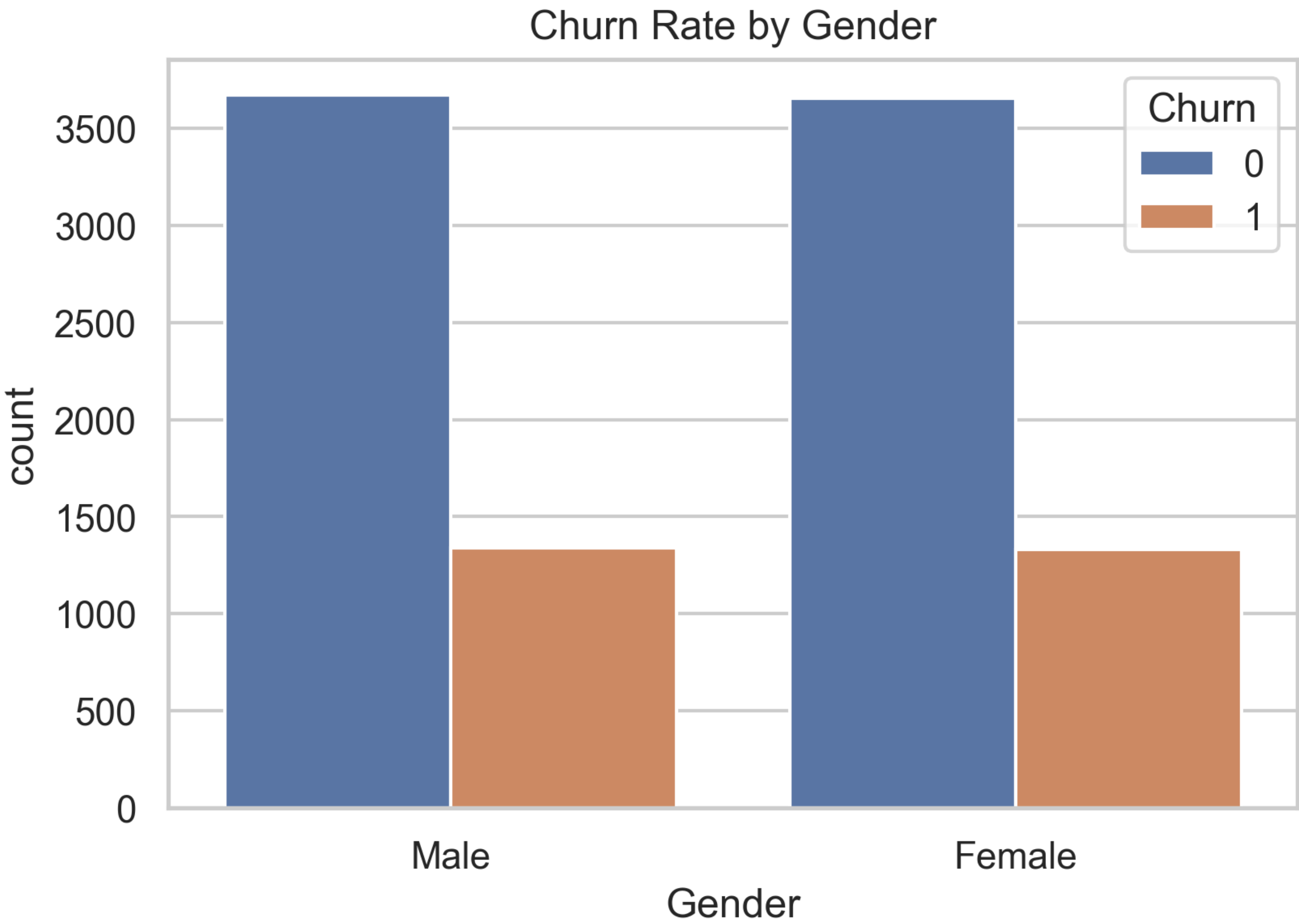


Metric	Value
Mean	$2,541.81
Median	$2,117.14

✅ Strongly dependent on **Tenure × Monthly Charges**.

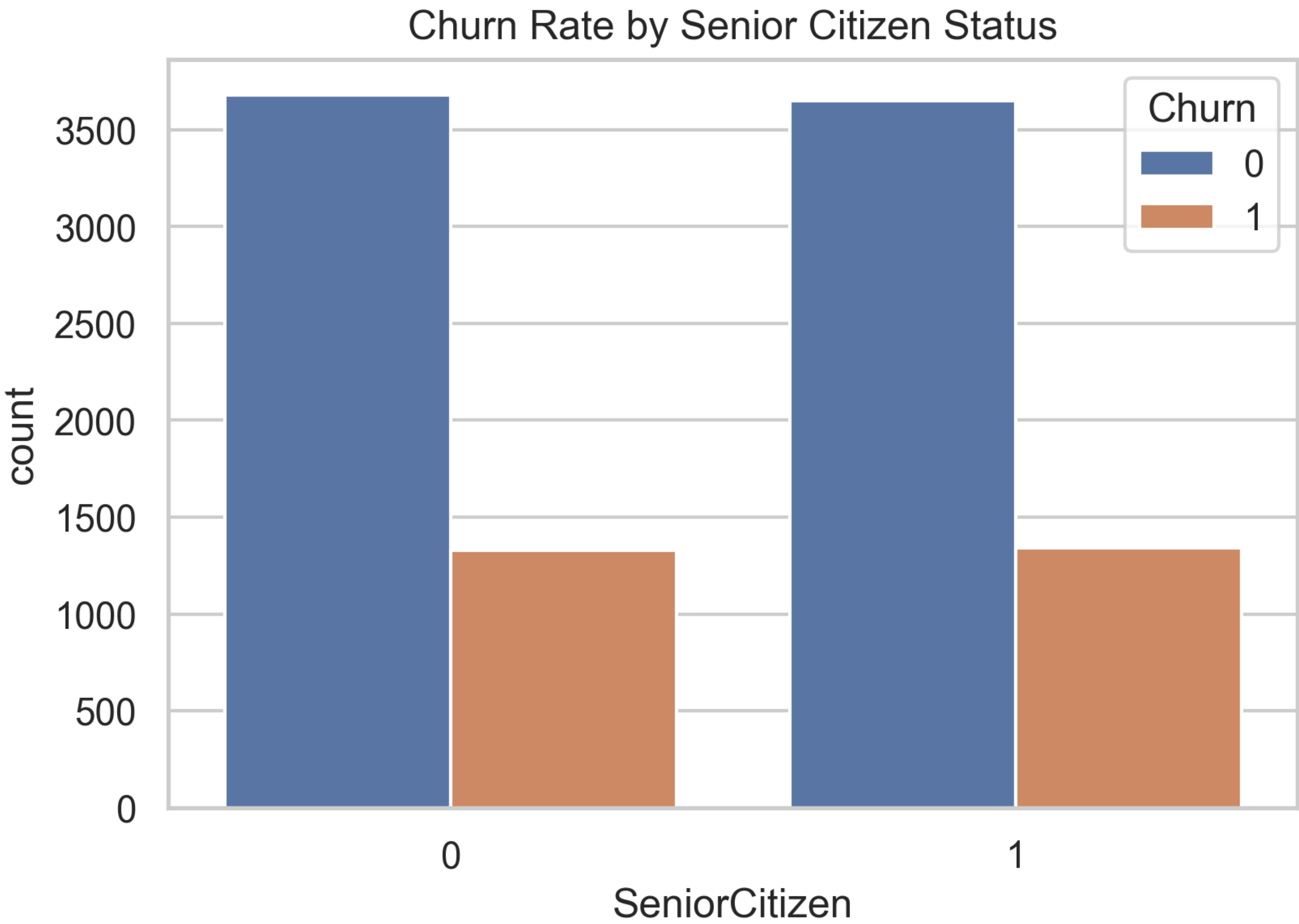
3.2 Bivariate Analysis

Gender vs Churn



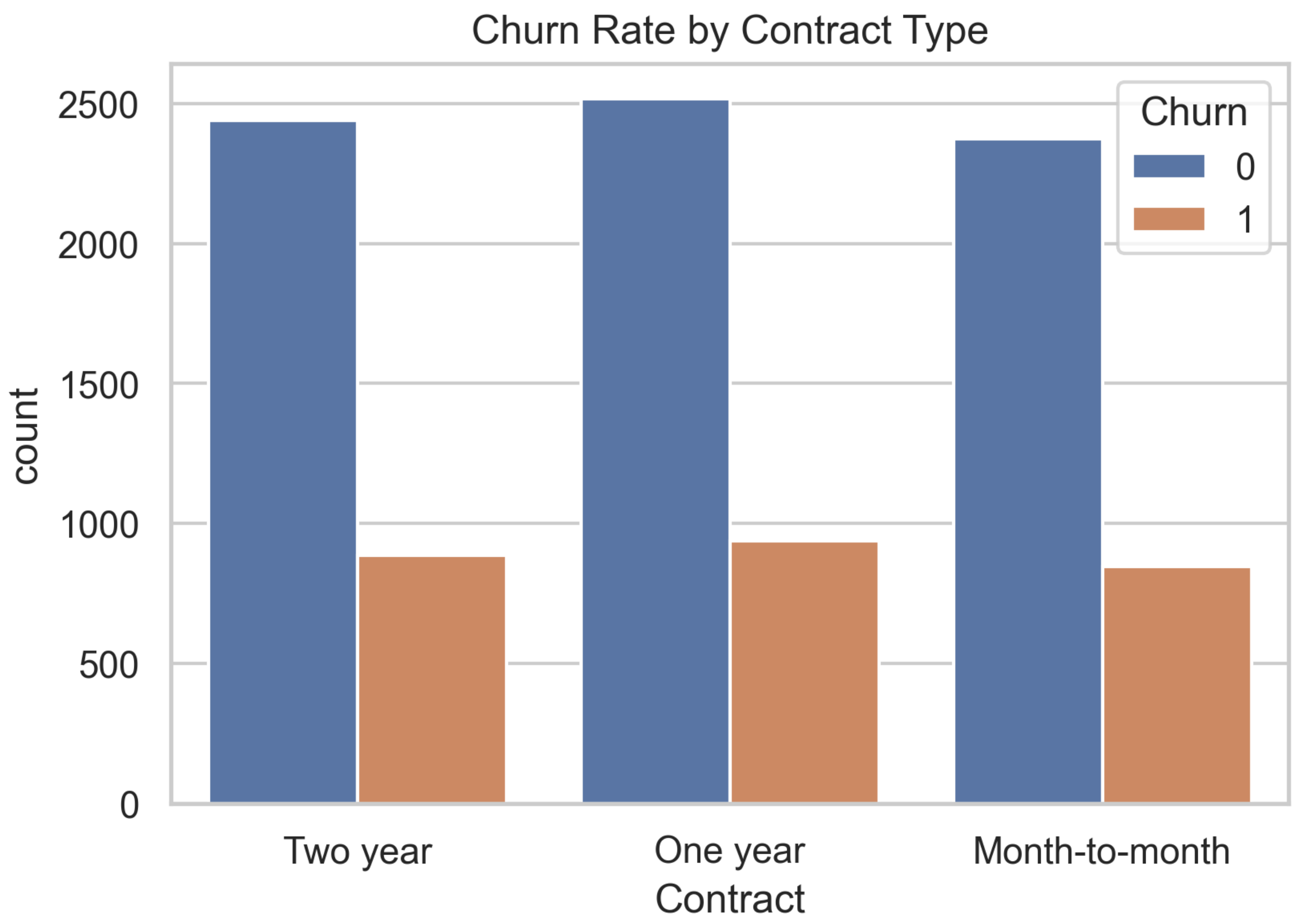
💡 No significant churn difference between genders.

Senior Citizen vs Churn



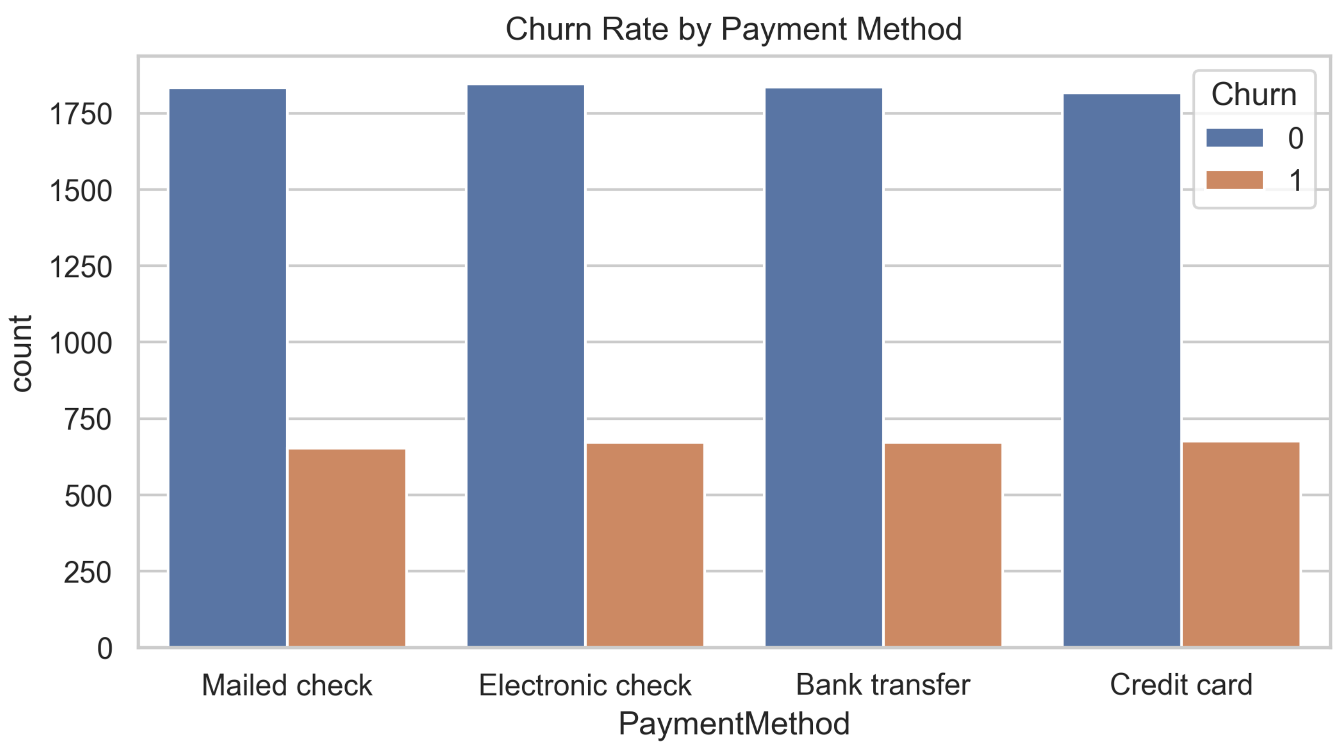
💡 Age has minimal churn impact.

Contract Type vs Churn



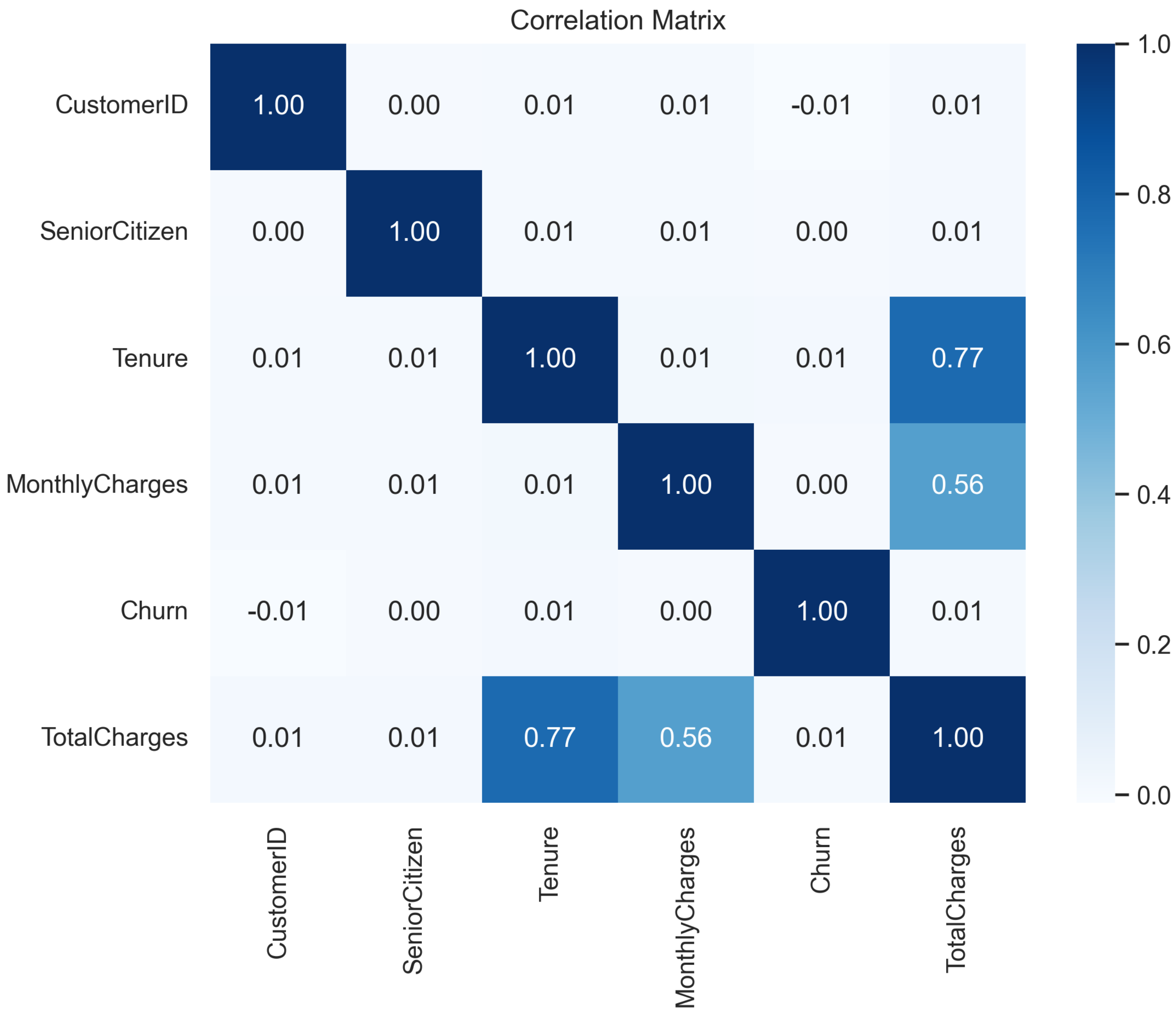
💡 Slightly higher churn in **1-year contracts**.

Payment Method vs Churn



💡 Churn is uniform across payment methods.

3.3 Correlation Insights



- **Total Charges ↔ Tenure:** 77% correlation
- **Total Charges ↔ Monthly Charges:** 56% correlation
  💡 Multicollinearity must be handled.

---

4. Feature Engineering

Key Steps

1. **Train/Test Split** – Stratified to preserve churn ratio.
2. **Data Cleaning** – Checked duplicates, data types, leakage.
3. **Derived Features** – Tenure groups, high monthly flag, autopay indicator, contract-payment combo.
4. **Encoding** – One-hot & K-fold target encoding.
5. **Scaling** – StandardScaler for numeric features.
6. **Multicollinearity Removal** – Dropped redundant features (`TotalCharges`, `AvgChargesPerMonth`).
7. **Feature Selection** – RandomForest + SelectFromModel confirmed **Contract**, **Tenure**, **MonthlyCharges** as top predictors.

---

5. Model Creation & Evaluation

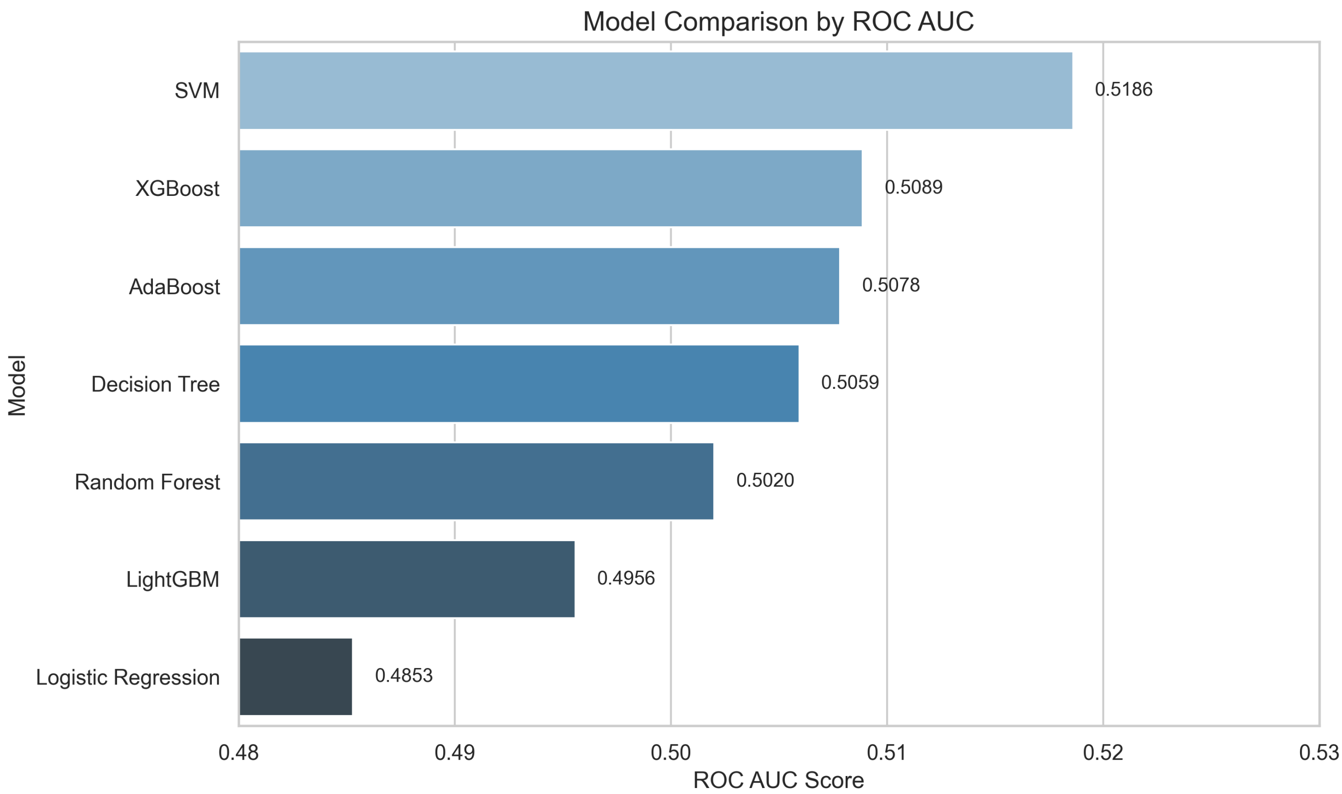
Models Tested

- Logistic Regression
- SVM
- Decision Tree
- Random Forest
- AdaBoost
- XGBoost
- LightGBM

Setup

- **Split:** 80/20 stratified
- **Tuning:** GridSearchCV (3-fold CV)
- **Metric:** ROC AUC

Performance Summary



Model	ROC AUC	Best Params
SVM	**0.5186**	`C=1, kernel='linear'`
XGBoost	0.5089	`n_estimators=200, max_depth=7`
AdaBoost	0.5078	`n_estimators=100`
Decision Tree	0.5059	`max_depth=3`
Random Forest	0.5020	`max_depth=5, n_estimators=200`

Model	ROC AUC	Best Params
Logistic Regression	0.4853	`C=10, penalty='l2'`
LightGBM	0.4956	`n_estimators=100, max_depth=5`

---

6. Key Observations & Next Steps

- All models performed **close to random guess (0.5 ROC AUC)**.
- SVM performed best but still insufficient for deployment.
- **Likely Issues:**
  - Features lack strong churn signal
  - Possible data noise or hidden imbalance
  - Need domain-specific feature engineering

Next Actions:

- Gather behavioral & usage pattern data
- Explore advanced interaction features
- Try SMOTE/class weighting if imbalance impacts learning

✅ Conclusion

The current dataset and features are **not sufficient** for accurate churn prediction.
The **SVM (C=1, linear kernel)** achieved the highest ROC AUC (**0.5186**) but requires significant feature and data enhancement before production use.
