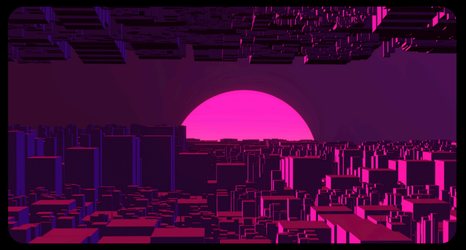
Machine Learning I – Final Project
BIA-5302-0LB

Telco Customer Churn Prediction using Machine Learning

A Predictive Analytics Approach
Using Python & Scikit-learn

Presented by : Karan Ethirajulu
Maheshwaran Sivakumar
Medline Jose
Divya Markose

Project Description of Operations



Objective:
To identify and predict customer churn in the telecommunications sector using historical customer data and machine learning models.

Data Cleaning & Preprocessing

Preparing raw data by handling missing values, outliers, and inconsistencies for analysis

Feature Engineering

Creating or modifying input variables to improve model performance and predictive accuracy.

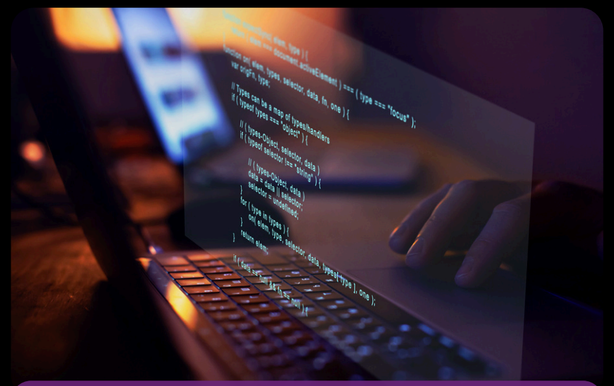
Model Development (Classification)

Building ML algorithms to categorize data into predefined classes or labels.

Performance Evaluation & Visualization

Assessing model effectiveness using metrics and graphical representations for insights.

Project Objectives



Predict Churn Likelihood

Forecast customer attrition risk using historical data and ML classification models.

Analyze Key Churn Drivers

Identify influential factors (e.g., usage patterns, demographics) causing customer churn.

Compare Classification Models

Evaluate algorithms (e.g., Logistic Regression, Random Forest) to select the best performer.

Reduce Churn Strategically

Recommend retention tactics (e.g., discounts, engagement) based on model insights.

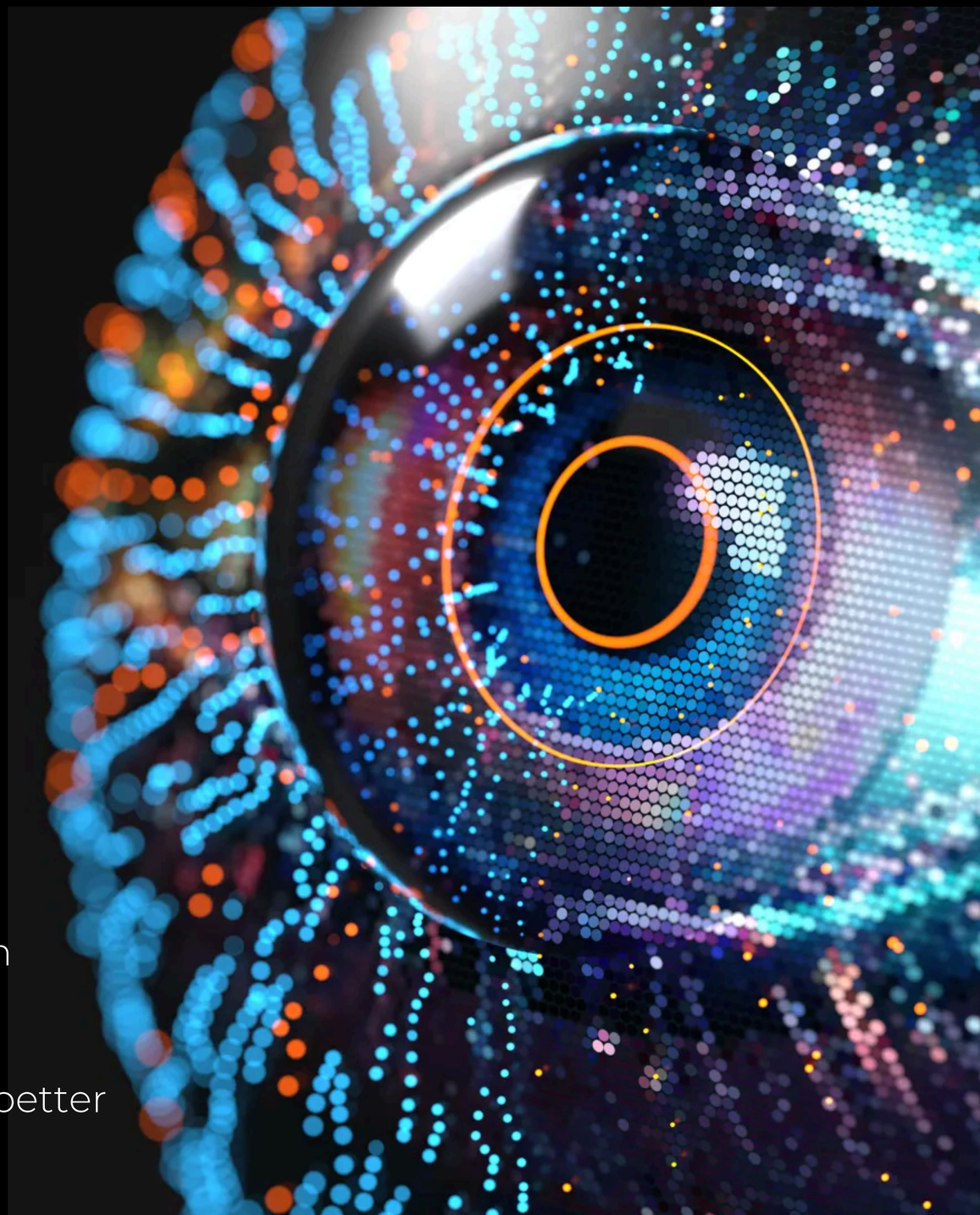
Industry Importance

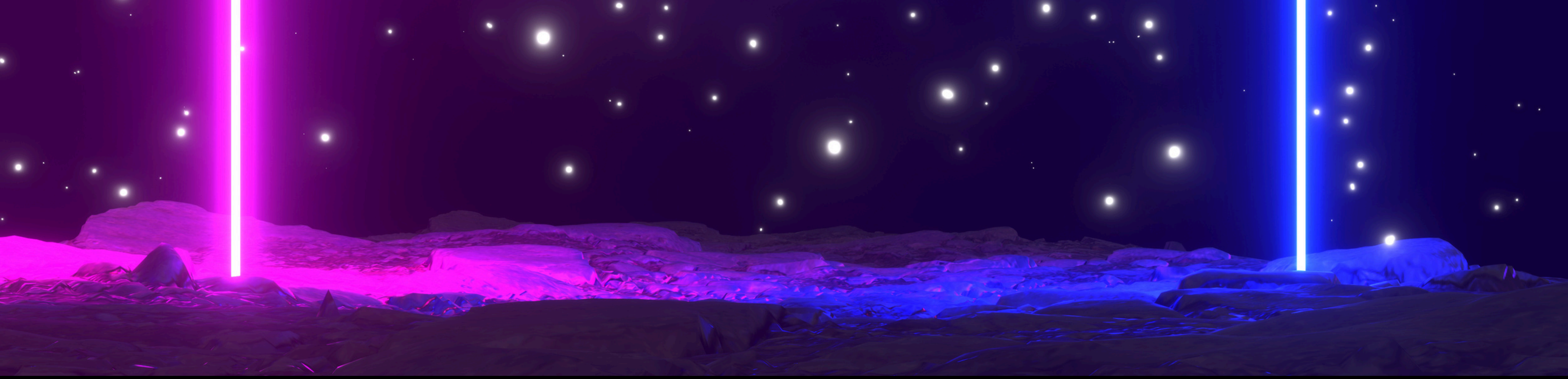
WHY THIS MATTERS:

- Churn directly impacts revenue and customer lifetime value.
- Predictive analytics help in proactive customer retention strategies.
- Telecom is a highly competitive industry; customer loyalty is key.

INDUSTRY RELEVANCE:

- Improves customer experience and reduces acquisition costs.
- Aids in targeting vulnerable customer segments with better offers





Dataset Overview

Dataset Used:

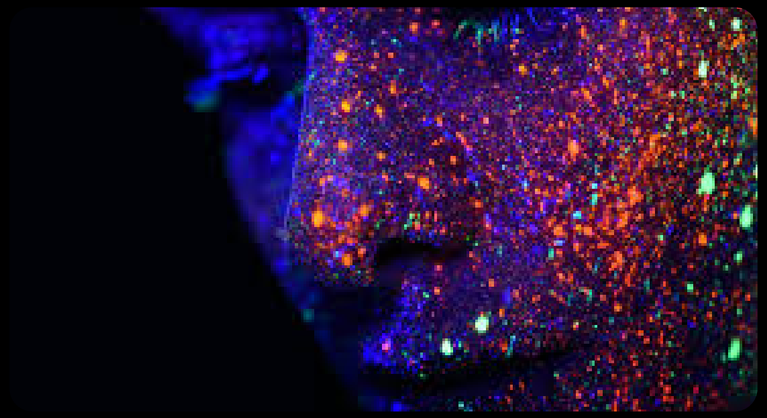
Telco Customer Churn Dataset (Publicly available via IBM)

Data Type:

Secondary data

Dataset Highlights:

Customer churn data includes service subscriptions, account details (tenure, billing), demographics (gender, age), and a binary churn indicator for attrition analysis.



Data Cleaning & Preprocessing

HANDLED MISSING VALUES 01

Converted TotalCharges to numeric and filled missing values with the median.

REMOVED IRRELEVANT COLUMNS 02

Dropped customerID.

ENCODED CATEGORICAL VARIABLES 03

Binary Encoding: Yes/No → 1/0
Gender: Male → 1, Female → 0
One-hot encoding: Contract, PaymentMethod, TenureGroup, etc.

OUTLIER TREATMENT 04

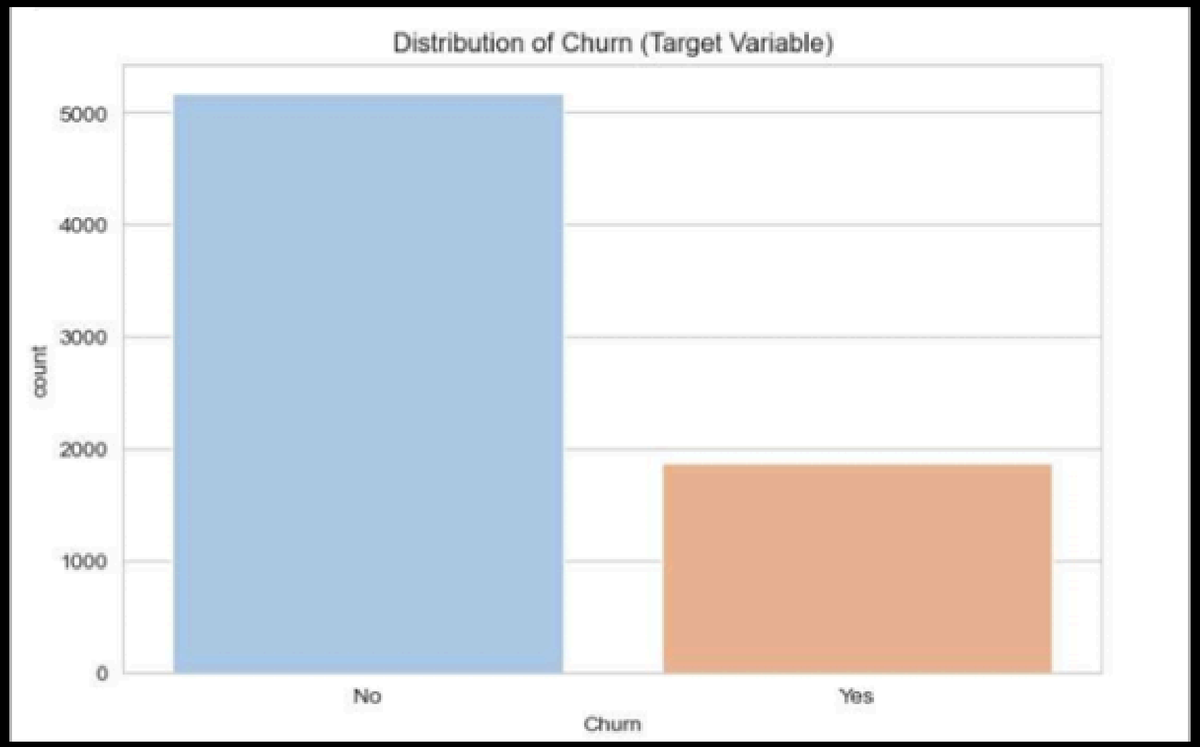
IQR method applied on numerical columns.

Data Visualization

Description:

This bar chart illustrates the distribution of the target variable Churn.

- Around 26.5% of customers have churned (Yes) while 73.5% have not (No).
- This highlights a class imbalance, which is important to address during model evaluation to avoid biased predictions



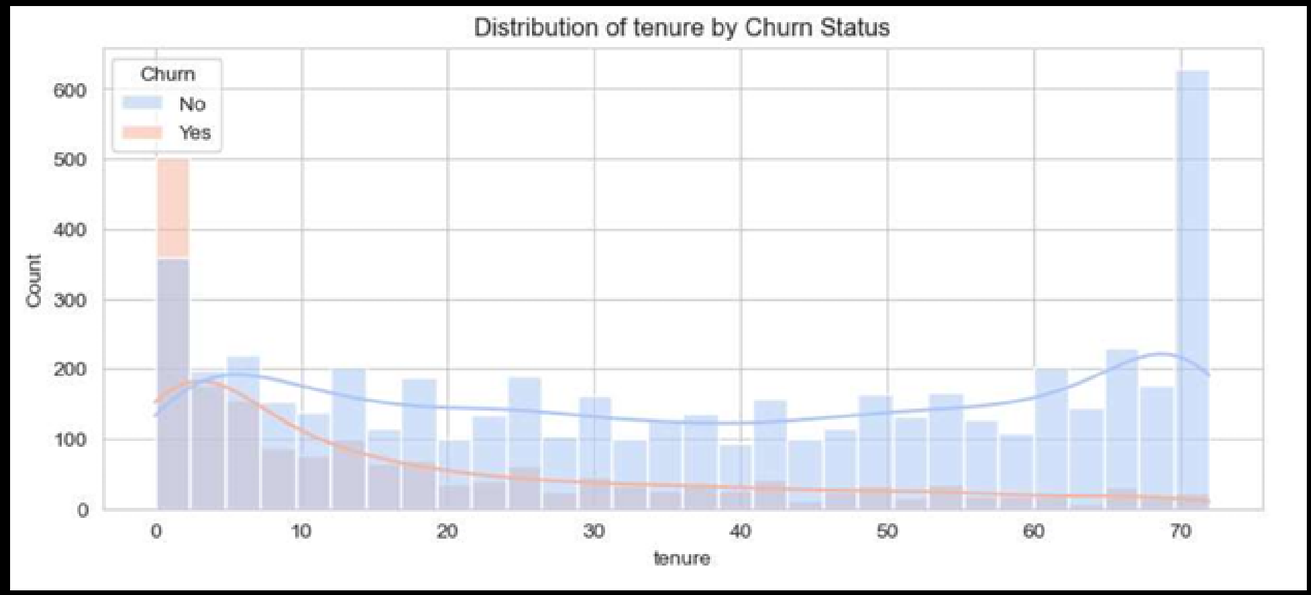
Data Visualization

Description:

a) Tenure Distribution

Insight:

- Customers with lower tenure (less than 12 months) are significantly more likely to churn.
- Long-term customers show higher retention, suggesting tenure is a key predictor.



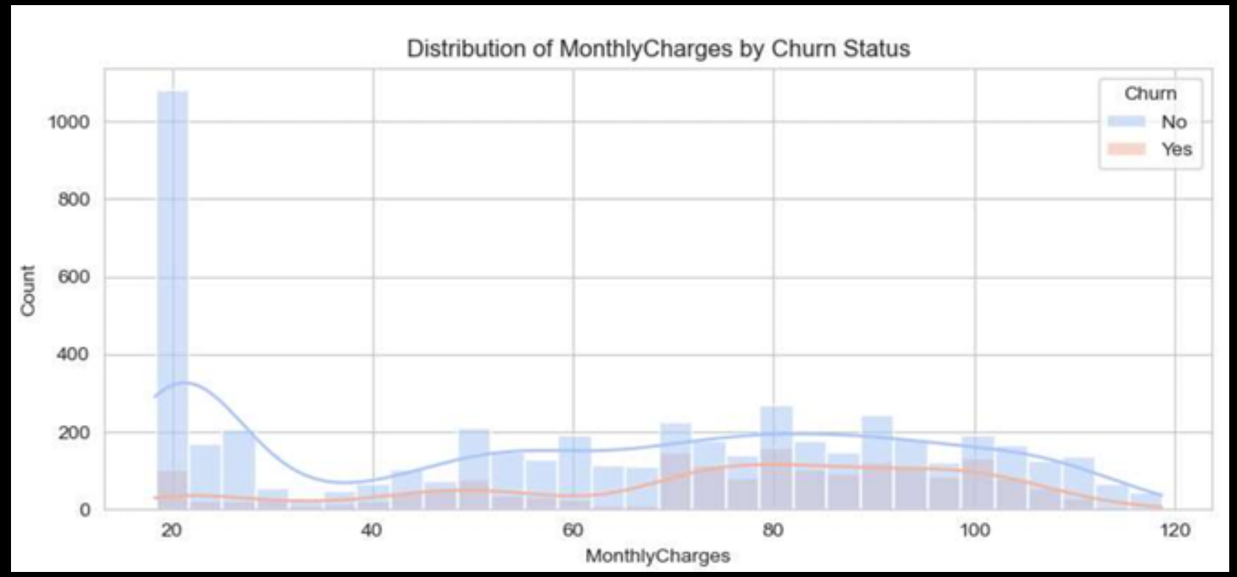
Data Visualization

Description:

b) Monthly Charges Distribution

Insight:

- Higher monthly charges are associated with churn.
- Churned customers tend to be concentrated in the $70–$90 range.



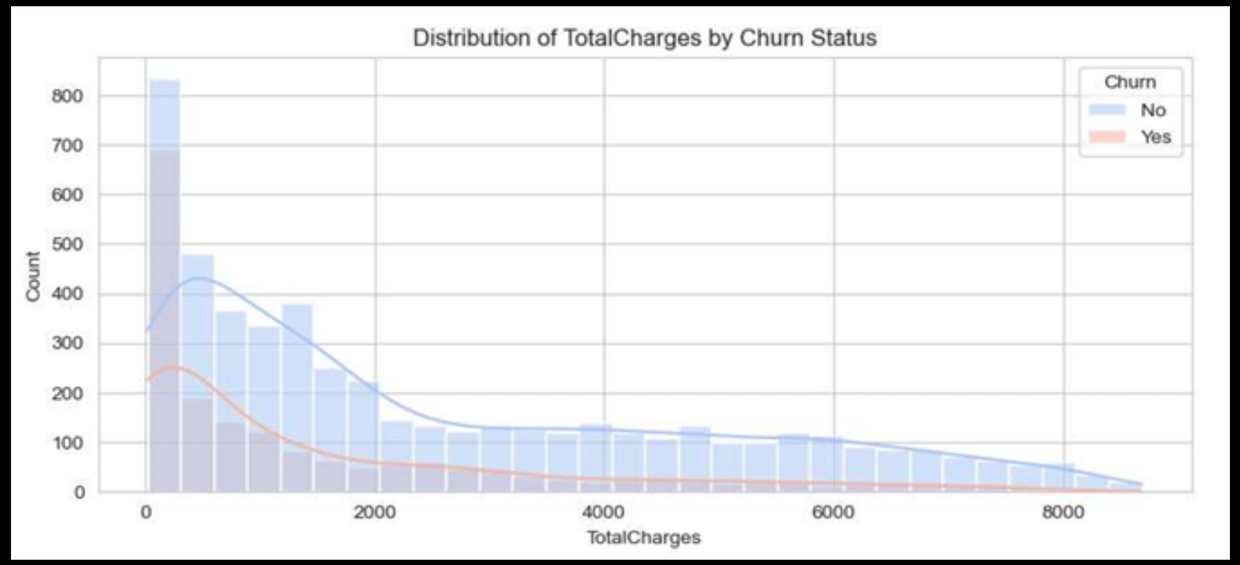
Data Visualization

Description:

c) Total Charges Distribution

Insight:

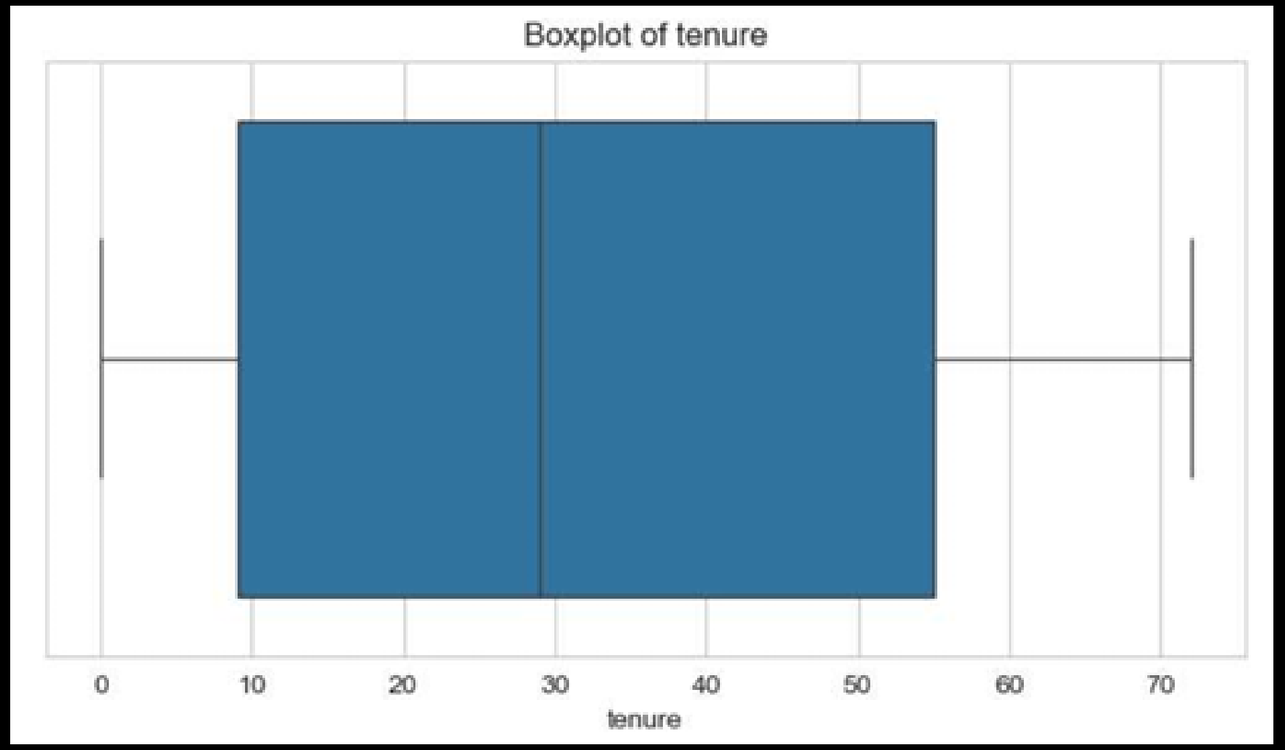
- Customers with lower total charges are more likely to churn, which aligns with their shorter tenure.
- As total charges increase, churn decreases.



Data Visualization

a) Boxplot for tenure

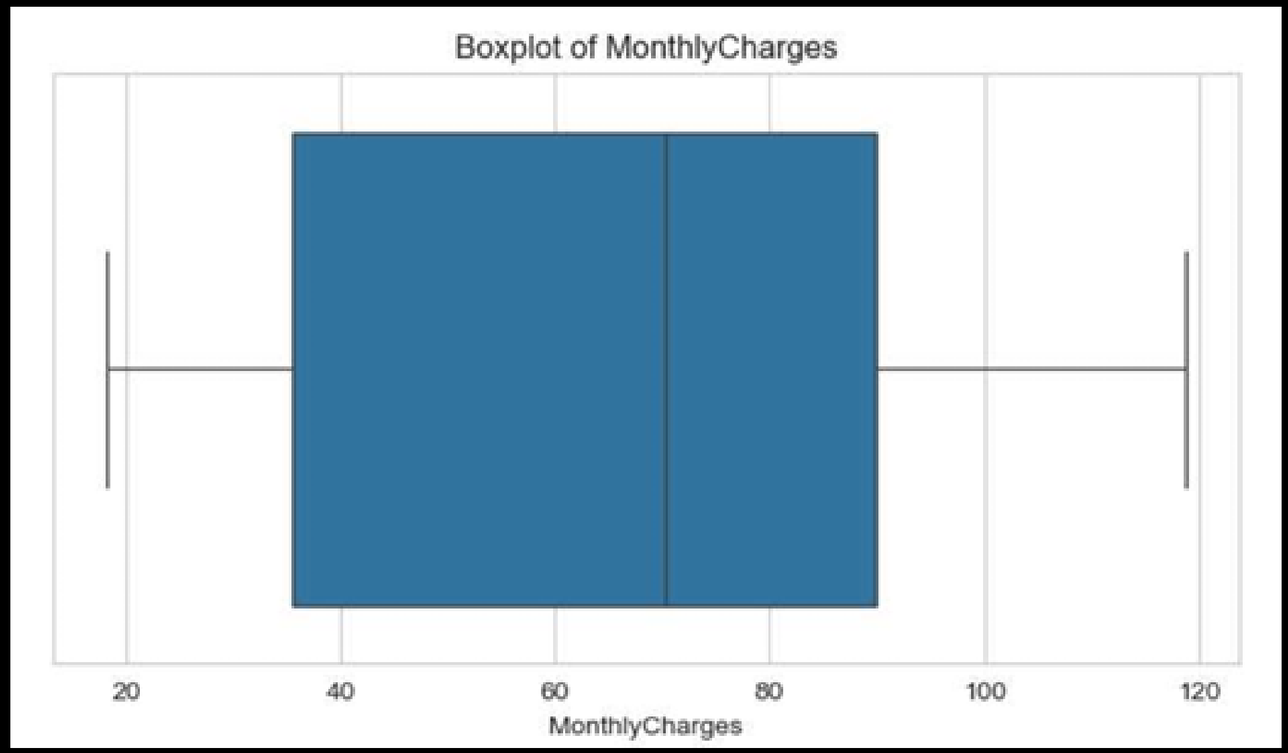
- Insight 1: The distribution of tenure is positively skewed, indicating most customers are relatively new.
- Insight 2: Very few extreme values, meaning customer duration is mostly consistent across the dataset.
- Business Insight: Short tenure customers are more at risk for churn; long-tenured ones tend to stay.



Data Visualization

b) Boxplot for MonthlyCharges

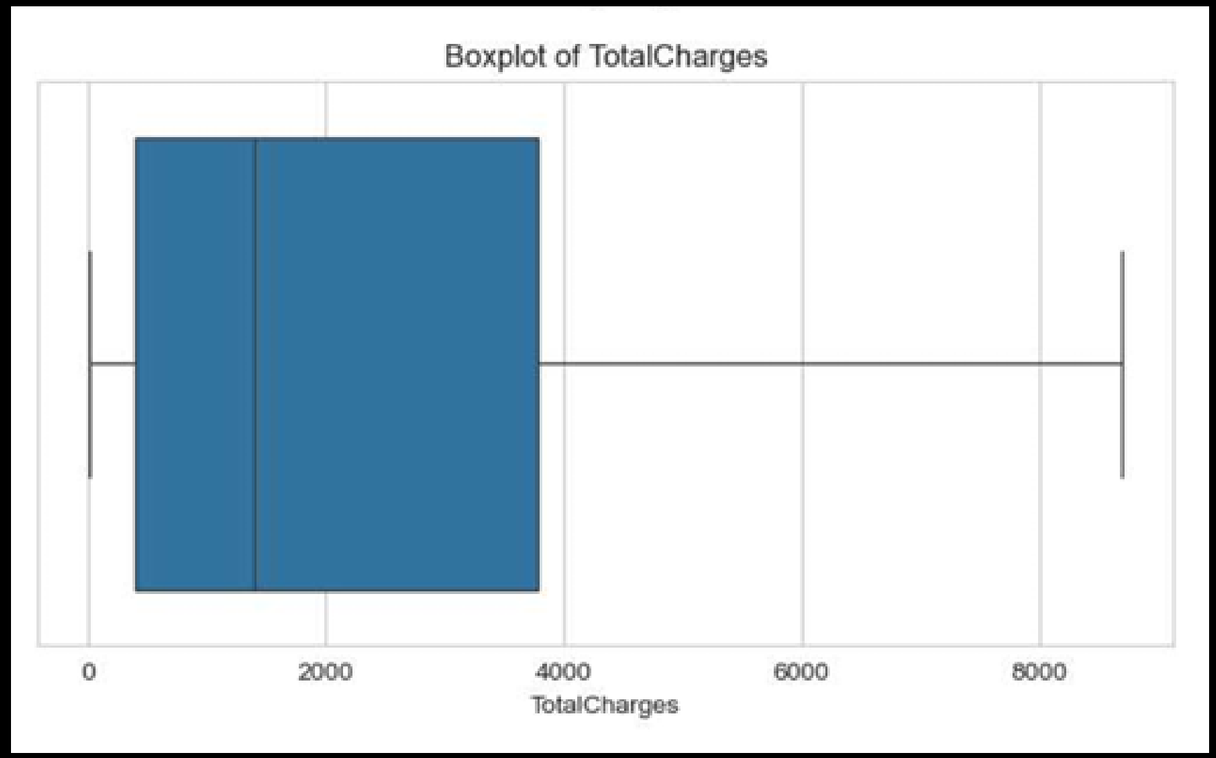
- Insight 1: There are a few high-end outliers where customers pay significantly more than average.
- Insight 2: The median MonthlyCharge is closer to the lower quartile, meaning a larger portion of customers pay less than average.
- Business Insight:
  - High-paying customers might churn hence considering bundles or discounts might be good option .



Data Visualization

c) Boxplot for TotalCharges

- Insight 1: Significant outliers observed at the higher end, representing customers who have spent a lot over time.
- Insight 2: These outliers correspond to long-term loyal customers, but their presence can skew model predictions.
- Handling Strategy:
  - Used IQR method to cap extreme values, ensuring stable model performance.



Data Visualization

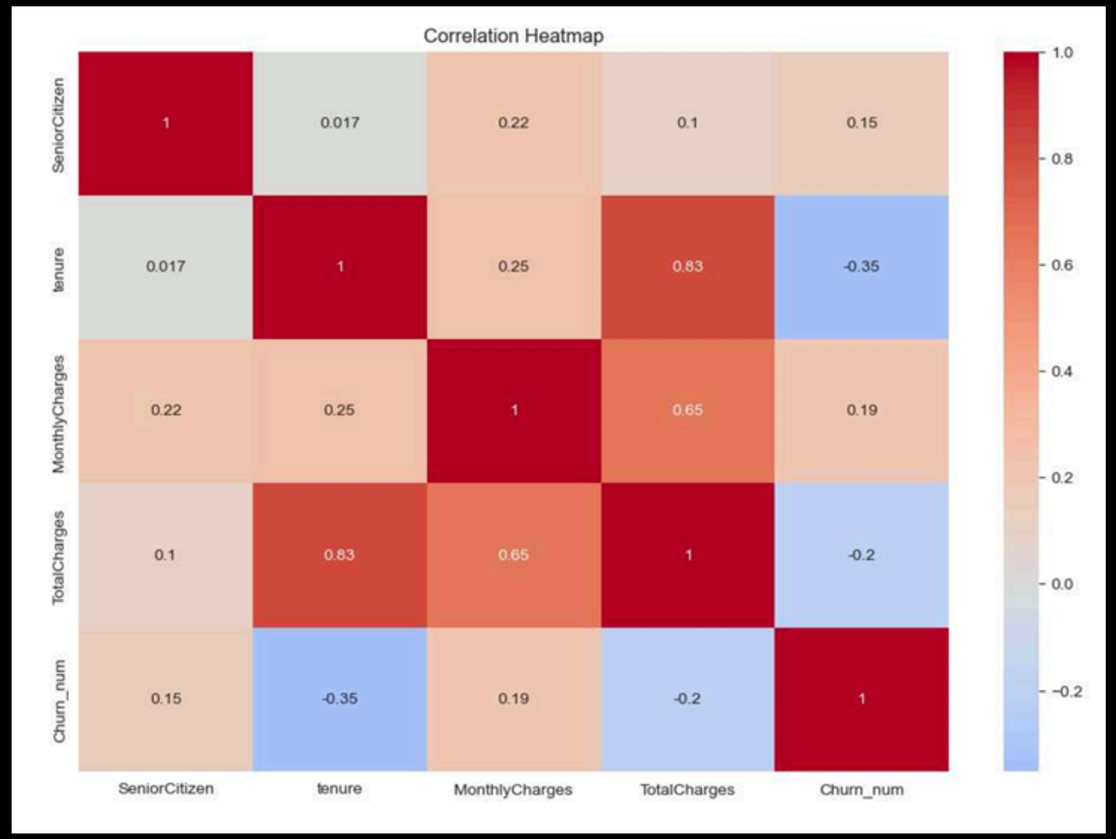
4. Correlation Heatmap

Top Correlations with Churn:

- Contract (month-to-month highly correlated with churn)
- Tenure (- correlation – the longer the tenure, the lower the churn)
- MonthlyCharges (+ correlation)

Insight:

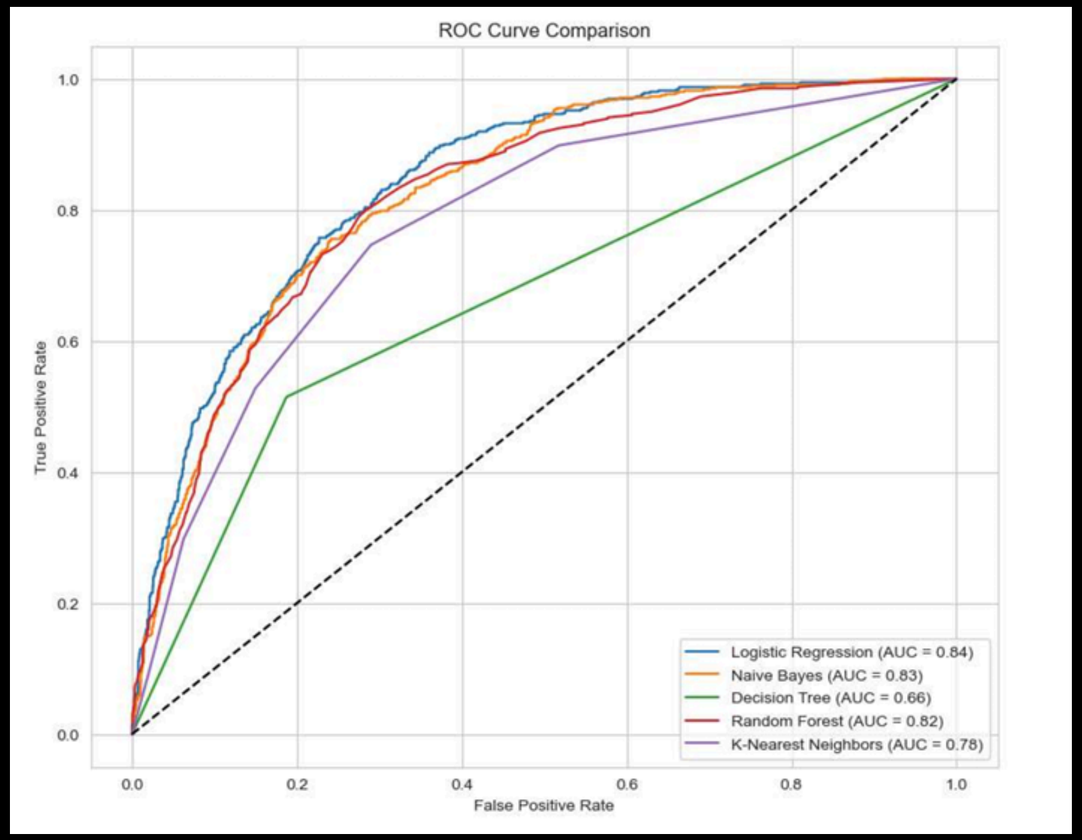
- Features with minimal correlation like gender or PhoneService were retained for completeness .Contract type and tenure are significant churn predictors.



Data Visualization

ROC Curve – Model Comparison

- ROC curves were plotted for 5 models: Logistic Regression, Random Forest, Decision Tree, Naive Bayes, and KNN.
- Random Forest had the highest AUC, followed closely by Logistic Regression.
- All models performed better than random guessing (diagonal line).



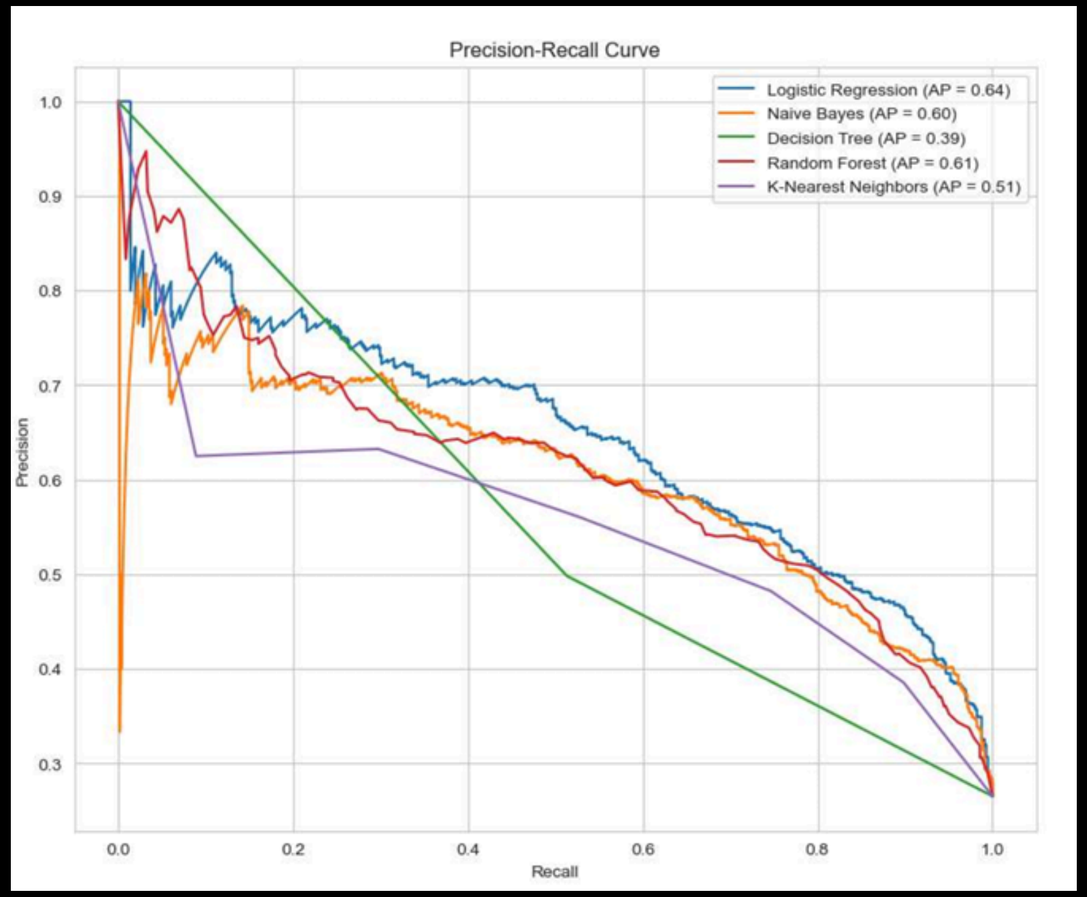
Data Visualization

7. Precision-Recall Curve

- Shows trade-off between precision and recall.
- Random Forest and Logistic Regression maintained high average precision scores.

Insight:

- Helps in choosing models based on business priority: catching more churners (recall) vs reducing false positives (precision).



Data Science Algorithms Used

Logistic Regression:

A linear model that estimates churn probability based on weighted feature relationships; serves as a simple, interpretable baseline.

Naive Bayes:

Applies Bayes' theorem assuming feature independence; performs fast churn prediction but less accurate due to feature correlations.

Decision Tree:

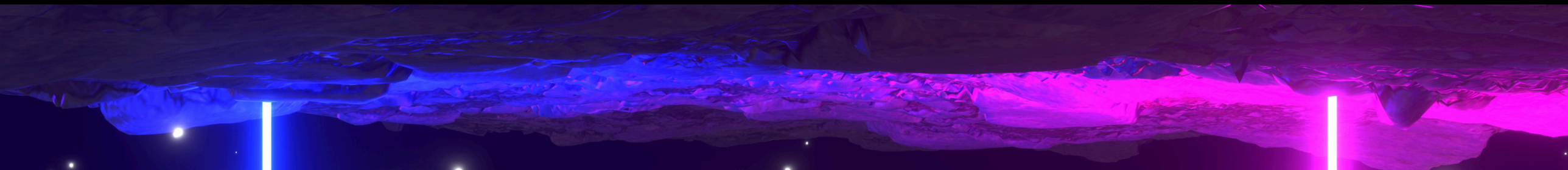
Splits data into branches using feature thresholds; captures nonlinear churn patterns but can overfit without pruning or regularization.

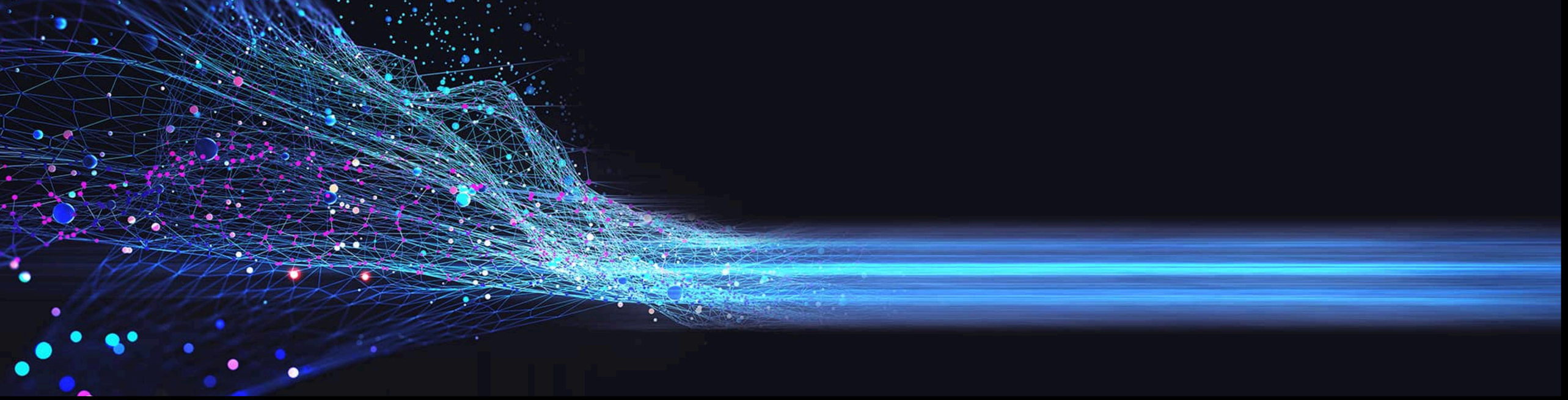
Random Forest:

An ensemble of decision trees that reduces overfitting; delivers higher churn prediction accuracy and feature importance rankings.

K-Nearest Neighbors:

Classifies churn by majority vote among closest customers; performance depends on scaling and optimal value of K.





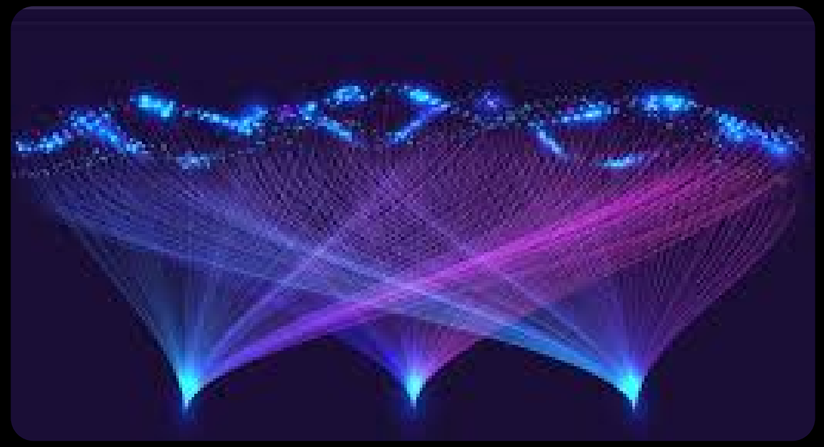
Developed Models

Model Training:

- Train-test split (70:30) with stratified sampling
- Feature scaling using StandardScaler

Models Built & Evaluated:

- Logistic Regression
- Naive Bayes
- Decision Tree
- Random Forest
- K-Nearest Neighbors



Model Evaluation & Results

Model	Accuracy	Precision (Churn)	Recall (Churn)	F1-Score (Churn)
Logistic Regression	80.0%	0.65	0.52	0.58
Random Forest	79.0%	0.64	0.49	0.55

Conclusion

- Predictive modeling is effective in identifying churn risk.
- Random Forest and Logistic Regression offered the best balance between performance and interpretability.
- Contract type and tenure are major predictors.
- Businesses can now proactively target likely churners with retention strategies.

Recommendations

- Offer discounts or loyalty rewards to customers on month-to-month contracts.
- Focus on providing reliable tech support and bundled services.
- Use model predictions to create a churn dashboard for marketing teams.