Customer Churn Prediction using Machine Learning

Report

1. Project Overview

The project aims at developing a machine learning-powered churn prediction system of a subscription business.

Customer churn is the process of customers ceasing to use a service of a given company- has a direct negative effect on the revenue, the costs of marketing and the growth of business in the long term. It is usually cheaper to retain than to attract new customers, and hence churn prediction is a crucial business objective.

Through the Telco Customer Churn Dataset, this project will precisely forecast the kind of customers who are likely to churn, the reasons behind the same and offer data-driven approaches to reduce churn.

The project is based on an end-to-end data science pipeline, which involves:

- Exploration and preprocessing of data.
- Selection and feature engineering.
- Model training, optimization and evaluation.
- Application and implementation.
- Practical business advice.

---

2. Problem Statement and Business Objectives

Problem Statement

The problem of customer churn is a significant issue to service-based business models like telecommunications, SaaS, and streaming services.
The objective is to construct a predictive model which can help to see the customers who are prone to churn before they do, so that the business can take proactive actions such as a specific offer or improvement in services.

Business Objectives:

- Create effective churn prediction model based on past customer history.
- Determine major churn causes including type of contract, tenure or mode of payment.
- Empower marketing and retention teams with data.
- Create a deployable pipeline that will be able to process incoming customer data and identify possible churners.

Success Criteria

- Model Performance: 80% ROC-AUC and 70% recall on churn class.
- Business Impact: Assist in minimizing customer churn by 10-15% by specific interventions.

---

3. Dataset Description

The data set is comprised of 7,043 records, one of a representative of a customer, and categorical, numerical, and binary variables are mixed.
It contains the demographics of customers, account details, services subscribed to, and billing details.

Main Categories:

- Demographics: Gender, SeniorCitizen, Partner, Dependents.
- Account Data: Tenure (months in company), Contract type, Paperless billing, type of payment.
- InternetService, OnlineSecurity, TechSupport, StreamingTV, StreamingMovies.
- Billing Information: MonthlyCharges, TotalCharges.
- Target Variable: Churn (Yes = left the customer, No = retained customer)

Initial Observations:

- There are continuous and categorical attributes present in the dataset.

- TotalCharges had to be converted to object to numeric.
- Target variable (Churn) is skewed, where about 26.5 percent of churners.

---

4. Success Metrics

The project uses a combination of classification metrics to ensure model reliability and practical usefulness:

Metric	Purpose
Recall (Sensitivity)	Prioritize identifying actual churners, minimizing false negatives.
Precision	Evaluate how many predicted churners are correct.
F1-Score	Balanced metric for uneven class distributions.
ROC-AUC	Measure model's ability to distinguish between churners and non-churners.
Accuracy	Overall performance indicator, though secondary to recall in churn prediction.

The chosen success criteria prioritize recall and ROC-AUC, as missing potential churners can cost more than false alarms in a business setting.

---

5. Exploratory Data Analysis

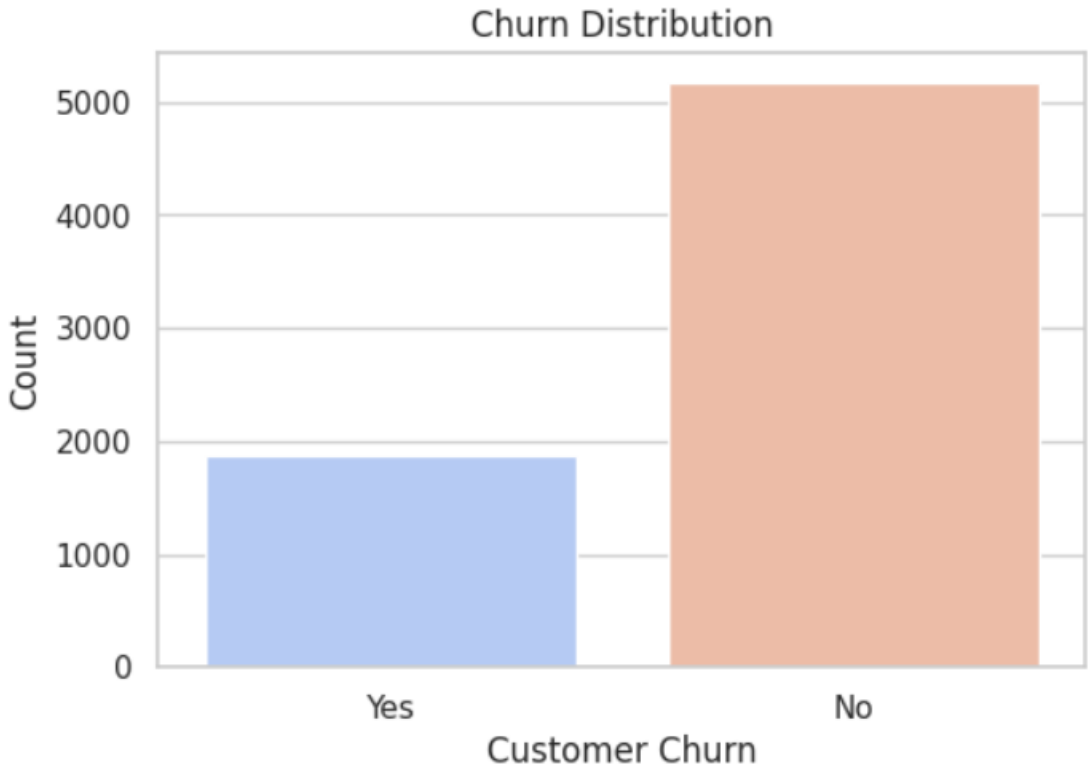
EDA was useful in uncovering trends, dependencies and correlations between variables that influence churn.

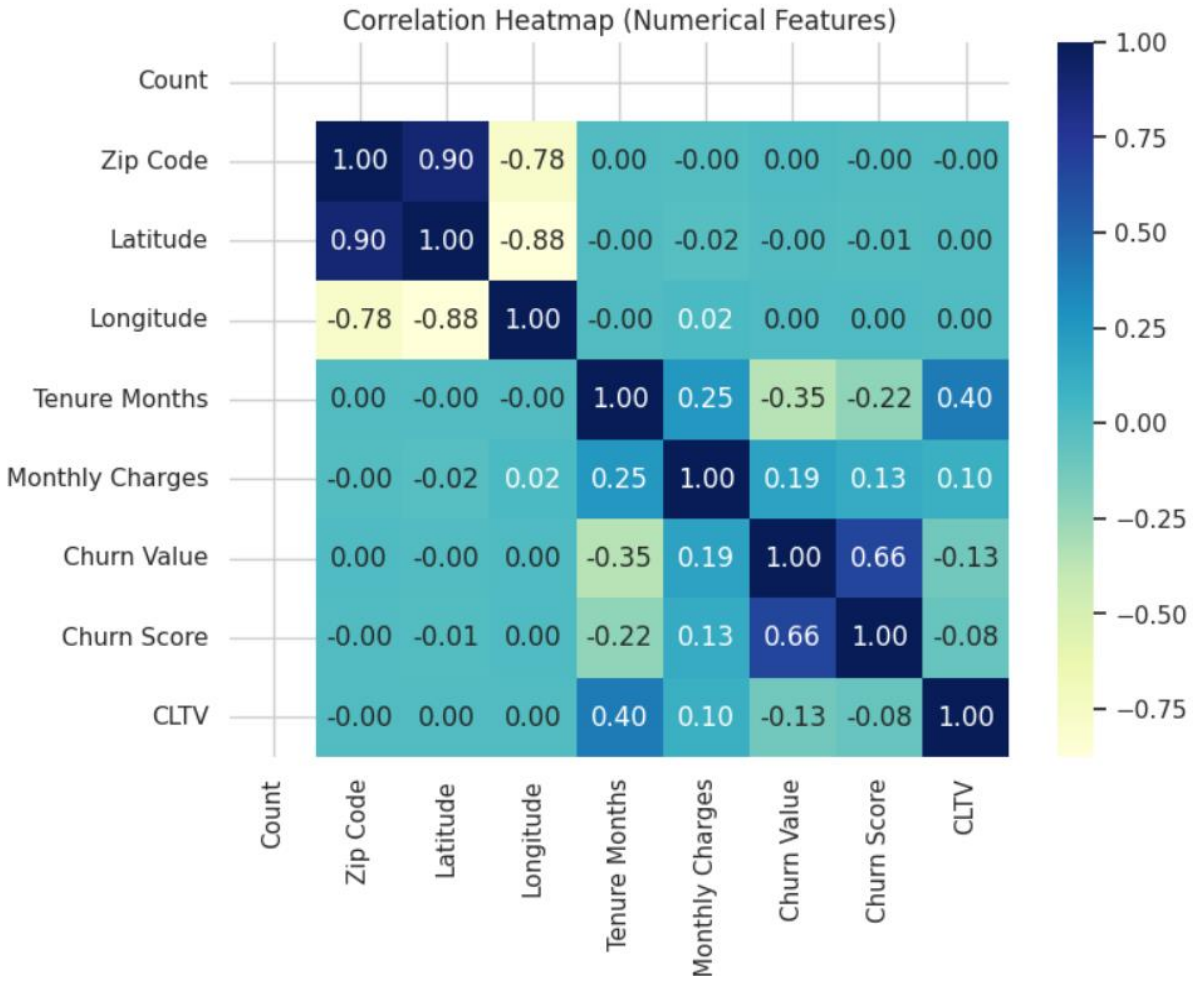
Key Insights:

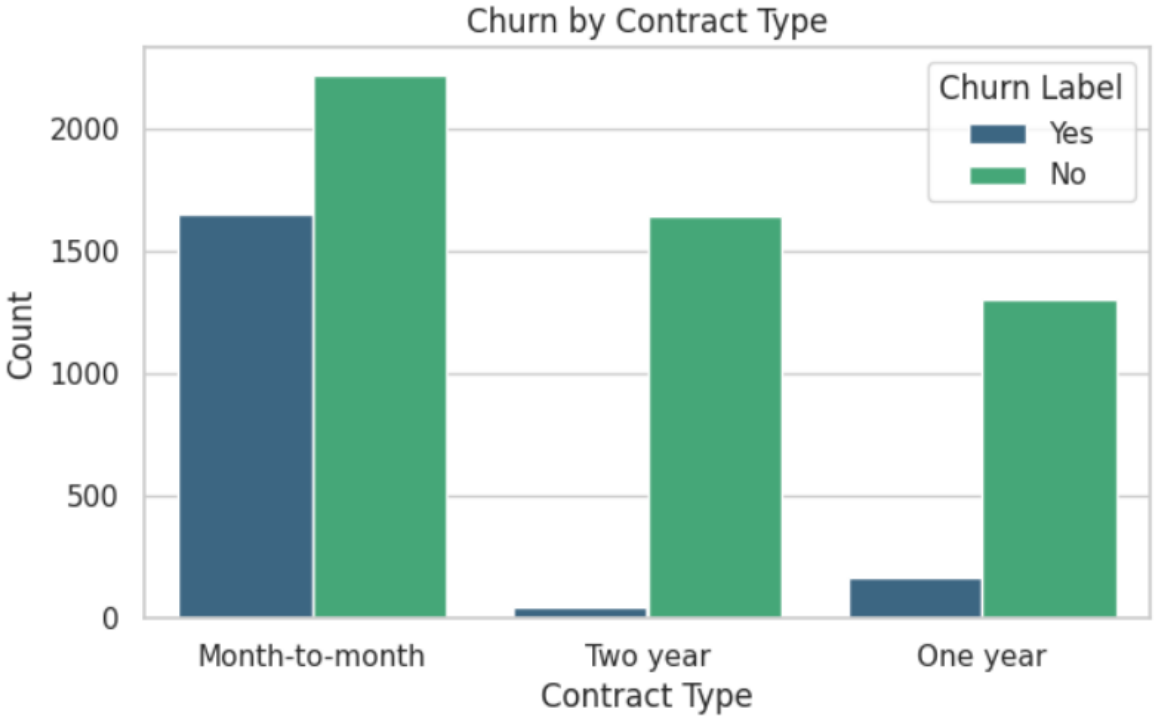
- Total Churn: 26.5% of the customers had churned, which showed that there was a moderate retention problem.
- Tenure Effect: The customers who had a tenure of less than 12 months had the greatest churn rates.
- Type of Contract: Customers with long-term contracts were more loyal whereas churn was much higher among the customers using month-to-month contracts.
- Method of payment: It was found that users who used electronic checks as payment method churned more than the users who used credit cards and automatic payments.
- Support Services: TechSupport or OnlineSecurity services were both strongly related with churn.
- Monthly Charges: High bills monthly were linked with high churn probability.

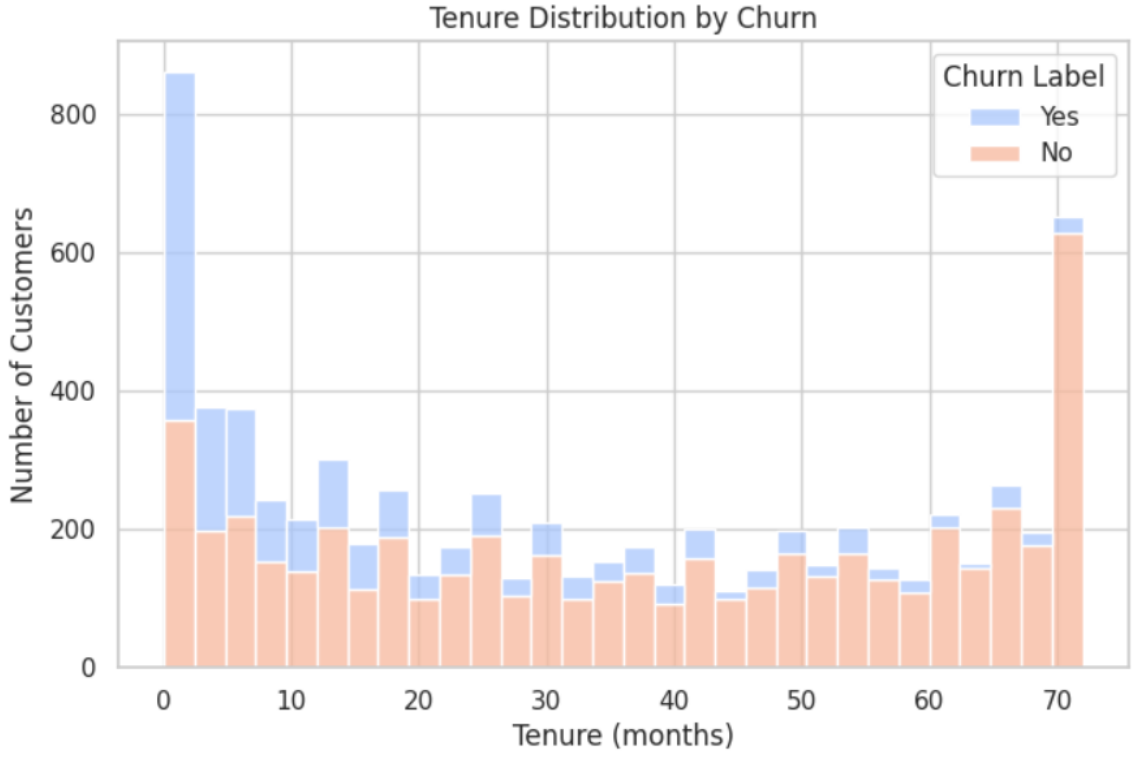
Visual Analysis Conducted:

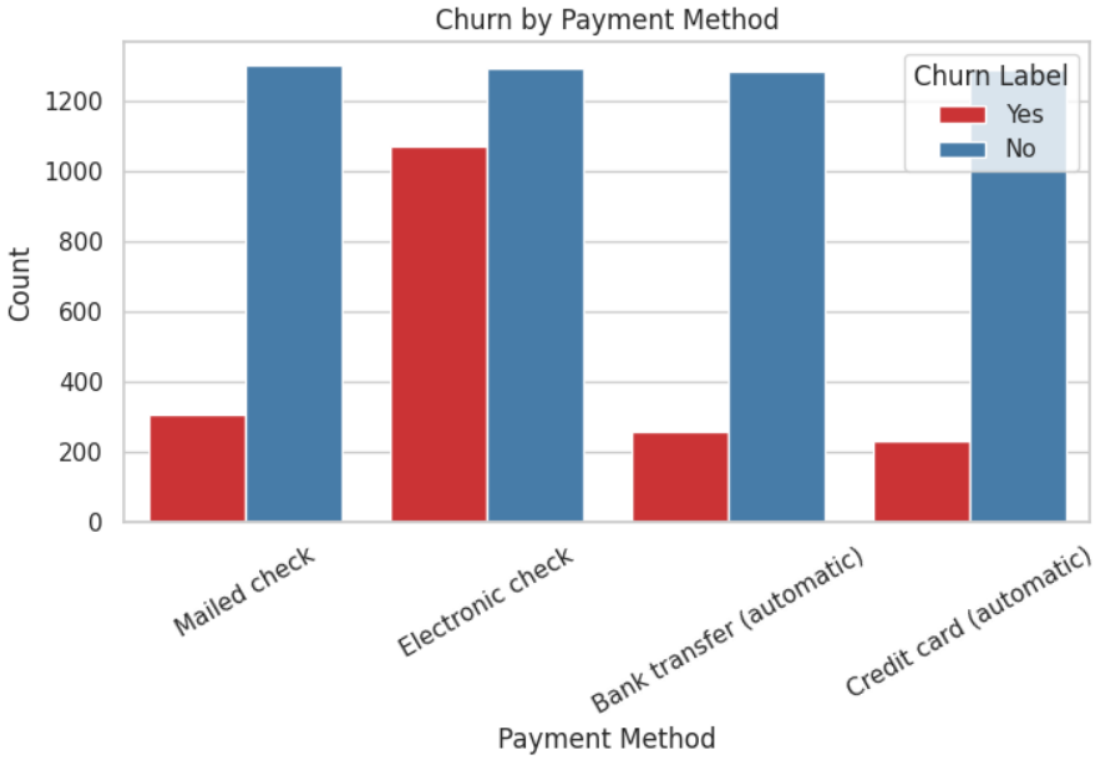
- Demographic churn distributions Count plots.
- Numerical feature correlation heatmap.
- Bar plots on churn by type of contract, tenure group and usage of services.
- These tests verified that the type of contract, payment method, tenure, and support services are some of the best churn indicators.


Churn Distribution
5000
4000
3000
2000
1000
0
Count
Yes
No
Customer Churn


Correlation Heatmap (Numerical Features)
Count
Zip Code
Latitude
Longitude
Tenure Months
Monthly Charges
Churn Value
Churn Score
CLTV
1.00 0.90 -0.78 0.00 -0.00 0.00 -0.00 -0.00
0.90 1.00 -0.88 -0.00 -0.02 -0.00 -0.01 0.00
-0.78 -0.88 1.00 -0.00 0.02 0.00 0.00 0.00
0.00 -0.00 -0.00 1.00 0.25 -0.35 -0.22 0.40
-0.00 -0.02 0.02 0.25 1.00 0.19 0.13 0.10
0.00 -0.00 0.00 -0.35 0.19 1.00 0.66 -0.13
-0.00 -0.01 0.00 -0.22 0.13 0.66 1.00 -0.08
-0.00 0.00 0.00 0.40 0.10 -0.13 -0.08 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Churn by Contract Type
Churn Label
Yes
No
Count
2000
1500
1000
500
0
Month-to-month
Two year
One year
Contract Type


Tenure Distribution by Churn
Churn Label
Yes
No
Number of Customers
800
600
400
200
0
0
10
20
30
40
50
60
70
Tenure (months)



---

6. Data Preprocessing

Purification and Metamorphosis:

- Conversed totalcharges of object to numeric; filled in missing values by median.
- Deleted redundant or invalid records.
- Inspection of handled outliers in monthly charges.

Encoding and Scaling:

- Binary Encoding: Yes/No features were changed to 1/0.
- One-Hot Encoding: This is used with multi-category variables such as InternetService and Contract.
- standardization: rescaling of numerical (tenure, MonthlyCharges, TotalCharges) variables using StandardScaler.

Feature Engineering:

- Established tenuregroup (0-12, 12-24, etc.) to mean customer loyalty categories.
- New numservices, the number of services active per customer.

Balancing:

- Applied SMOTE (Synthetic Minority Over-sampling Technique) to equalize the target variable to avoid bias in the majority class.
- The end dataset was a clean, standardized, balanced, and model training ready dataset.

---

7. Model Development

The models that were trained and assessed were as follows:

- Logistic Regression - interpretable model.
- Decision Tree Classifier -- can be easily explained, and can overfit.
- Random Forest Classifier - ensemble based model and good generalization.
- Gradient Boosting Classifier - serial model enhancing feeble learners.
- XGBoost Classifier optimized boosting with efficient learning.

Hyperparameter Tuning:

All models were optimized with the help of GridSearchCV and RandomizedSearchCV with 5-fold cross-validation.

Parameters tuned included:

- nestimators
- maxdepth
- minsamplessplit
- learning rate (to boost models)

---

8. Model Evaluation and Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.7711	0.5691	0.5722	0.5707	0.8145
Decision Tree	0.7548	0.5343	0.6043	0.5671	0.7116
Random Forest	0.7918	0.6052	0.6230	0.6140	0.8367
Gradient Boosting	0.7797	0.5702	0.6952	0.6265	0.8470
XGBoost	0.7832	0.5765	0.6952	0.6303	0.8438

Model Selection:

While Gradient Boosting and XGBoost achieved slightly higher ROC-AUC, Random Forest was selected as the final model for the following reasons:

- Delivered consistent and balanced performance across all metrics.
- Lower risk of overfitting compared to boosting models.
- Easier interpretability and feature importance extraction.

Final Model: Random Forest
Performance Summary:

- Accuracy: 79.18%
- Recall: 62.3%
- ROC-AUC: 83.67%

---

9. Feature Importance and Interpretation

Feature importance derived from the Random Forest model highlighted the most influential factors affecting churn:

Rank	Feature	Impact
1	Contract Type	Short-term contracts strongly linked with churn
2	Tenure	Shorter tenures show high churn likelihood
3	TotalCharges	Lower total spending often indicates new customers
4	MonthlyCharges	High bills can lead to dissatisfaction
5	TechSupport	Lack of support leads to higher churn
6	OnlineSecurity	Missing service linked to churn
7	PaymentMethod_ElectronicCheck	Payment inconvenience increases churn
8	InternetService_FiberOptic	Higher churn likely due to cost issues
9	PaperlessBilling	Associated with younger, more mobile users
10	SeniorCitizen	Older customers show moderate churn risk

Interpretation:

Customers on month-to-month contracts, with no additional support services, short tenure, and high monthly charges, are most likely to churn.
This indicates the importance of customer engagement early in their tenure and offering value-added services.

---

10. Deployment Pipeline

An API was developed based on Flask in order to operationalize the model.

Deployment Steps:

1. Image preprocessing pipeline and Random Forest model were trained, and then joblib was used to serialize them.
2. A Flask API has been created that has a /predict endpoint that accepts customer features inputed in the form of a JSON.
3. The output of API is the churn probability and classification (0 or 1)

Example Output:

```
{
  "probability": 0.72,
  "prediction": 1
}
```

Predictions above a threshold of 0.6 indicate a high likelihood of churn.

This system can be integrated into CRM dashboards or business portals for real-time monitoring.

---

11. Business Recommendations

According to the analysis and according to the main features that were identified and impacted churn, the following are some actionable business recommendations:

1. **Target Short-tenure Customers**: Customers with a shorter tenure are higher in the churn. Adopt specific interventions to attract and keep the new customers within the first few months. This could include:
- Improved Onboarding: Offer a good customer experience by availing the best services and resources in the initial months to make sure customers are already enjoy the service.
- Early Intervention: Keep an eye on the activity and satisfaction of new customers and be proactive to contact them in case they are either disengaged or unhappy.
- Welcome Offers/Discounts: Provide incentives or discounts to new customers in order to make them remain beyond the probation trading duration.
2. **Enhance Support Quality and Accessibility**: Tech support-related and online security were also found to be significant features. Accessibility and quality of customer service can have a strong influence on retention.

- 24/7 Support: Provide 24/7 support in many forms (phone, chat, email).
- Shorter Response Times: Improve the customer support process by making it shorter.
- Proactive Support: Predict the churning of the customers by making use of the churn prediction model and provide them with proactive support or check-ins.

3. **Encourage Long-term Contracts**: Month to month customers turn over at a higher rate. Customer retention can be enhanced by incentivizing customers to enter into long-lasting contracts (one-year or two-year) with the company.

- Discounted Rates: This will provide customers with good discounts by taking longer contracting.
- Value-Added Services: Offer extra services or features to those customers who make longer commitment.
- Effective Spoken Communication of Benefits: Be able to clearly spell out the cost savings and advantages of long-term contracts to the customers.

4. **Maximize Pricing and Billing**: Monthly billing and total billing are valuable ones.

- Review Pricing Tiers: Review and ensure price is competitive and seen as fair price of the services provided.
- Convenient Payment methods: Provide diverse payment methods.
- Transparent Billing: Ensuring that bills are easily understood and do not contain any surprises.

5. **Take Advantage of Electronic Payments**: Electronic check was singled out as a payment method that was linked with a greater churn.

- Support Automatic Payments: Support and reward automatic payment methods such as bank transfers or credit cards.
- Research Problem with Electronic Check: learn why electronic check customers may be churning higher and correct any problems.

6. **Personalize Offers and Communications**: Use the insights of the feature importance to segment the customers and retention offers and communications. As an example, offer relevant offers to target customers having high monthly charges or specific internet service type.

---

12. Learning Outcomes:

This project allowed me to experience the following in practice:

- In-depth cleaning and preprocessing of data.
- The imbalanced data addressed with the help of SMOTE.

- Enhancing models through ensemble and boosting models in classification.
- Measuring performance using a variety of statistical measures.
- Developing deployable ML systems.
- Turning model knowledge into business value and strategy.

---

13. Conclusion

The project met its aim of making a good prediction of customer churn that was highly interpretable.

The Random Forest model provided good performance with:

- Accuracy: 79.18%
- ROC-AUC: 83.67%
- Recall: 62.3%

Other critical churn drivers are type of contract, tenure, support services and billing patterns. The information in the model is used to enable organizations to identify potential risk customers at the initial stage and develop relevant retention campaigns to enhance customer satisfaction and long-term profitability.

---

14. Future Enhancements:

- Include churn tracking in business dashboards in real-time.
- Make decisions using SHAP explainability visualizations.
- Use uplift modeling to assess retention campaign effectiveness.
- Train the model again periodically using new customer data.
- Continuous learning should be automated.
