Machine Learning Model Development and Evaluation Summary

1. Dataset Overview and Preprocessing

1.1.Dataset Information: Use the IBM Telco Customer Churn dataset available on Kaggle.

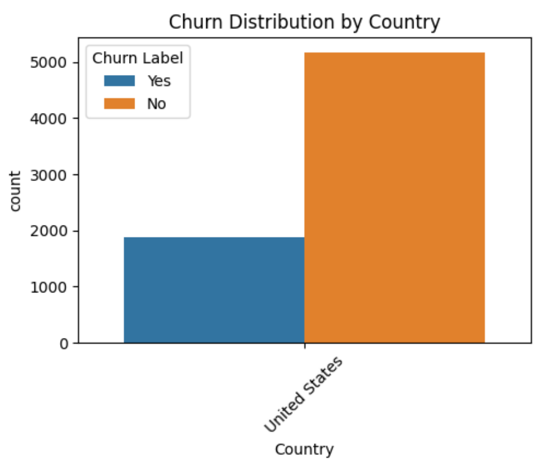
1. Content: Customer information, services, and churn-related details.
2. Features: Demographics, services used, contract details.
3. Target Variable: Churn (binary: 0 or 1).

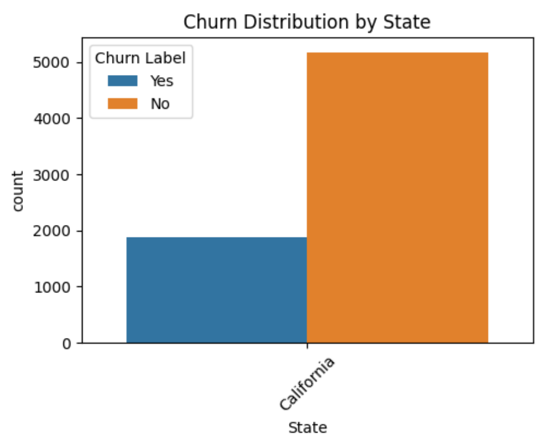
1.2.Preprocessing Steps:

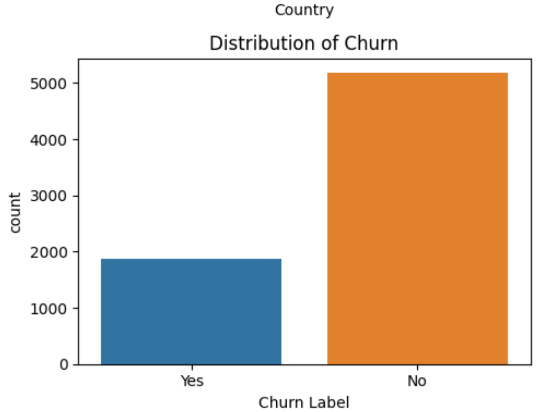
1. Handling Missing Values: Coerced 'Total Charges' to numeric, filling NaN values based on 'Tenure Months' and 'Monthly Charges.'
2. Data Cleaning: Checked and handled duplicates.
3. Feature Engineering: Created new features like 'Tenure Years' and 'Monthly to Total Charges Ratio.'
4. Categorical Data Handling: One-hot encoding, label encoding, and value replacements.

2. Exploratory Data Analysis (EDA)

2.1. Geographic, Demographic, and Service Utilization Analysis







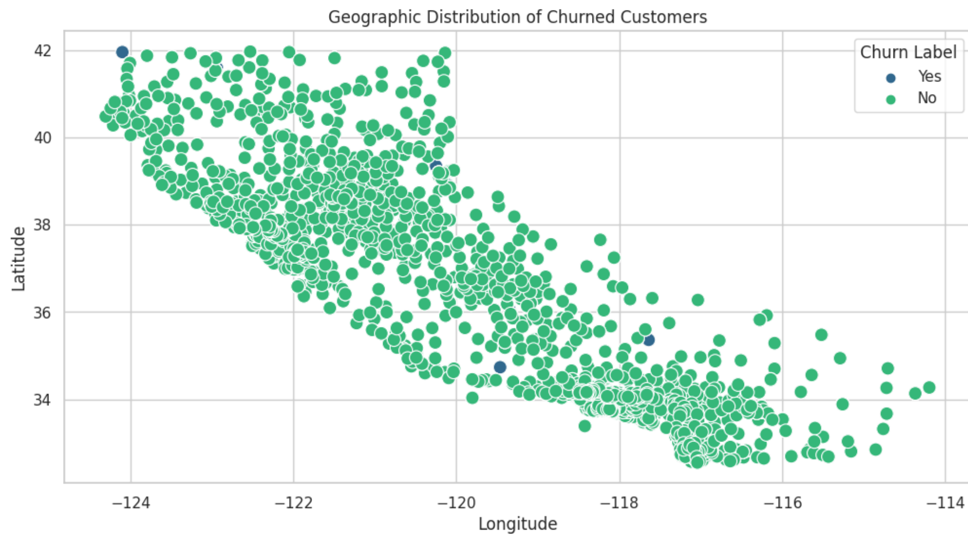


Fig: Explored distribution of churn across various Geographic categories.

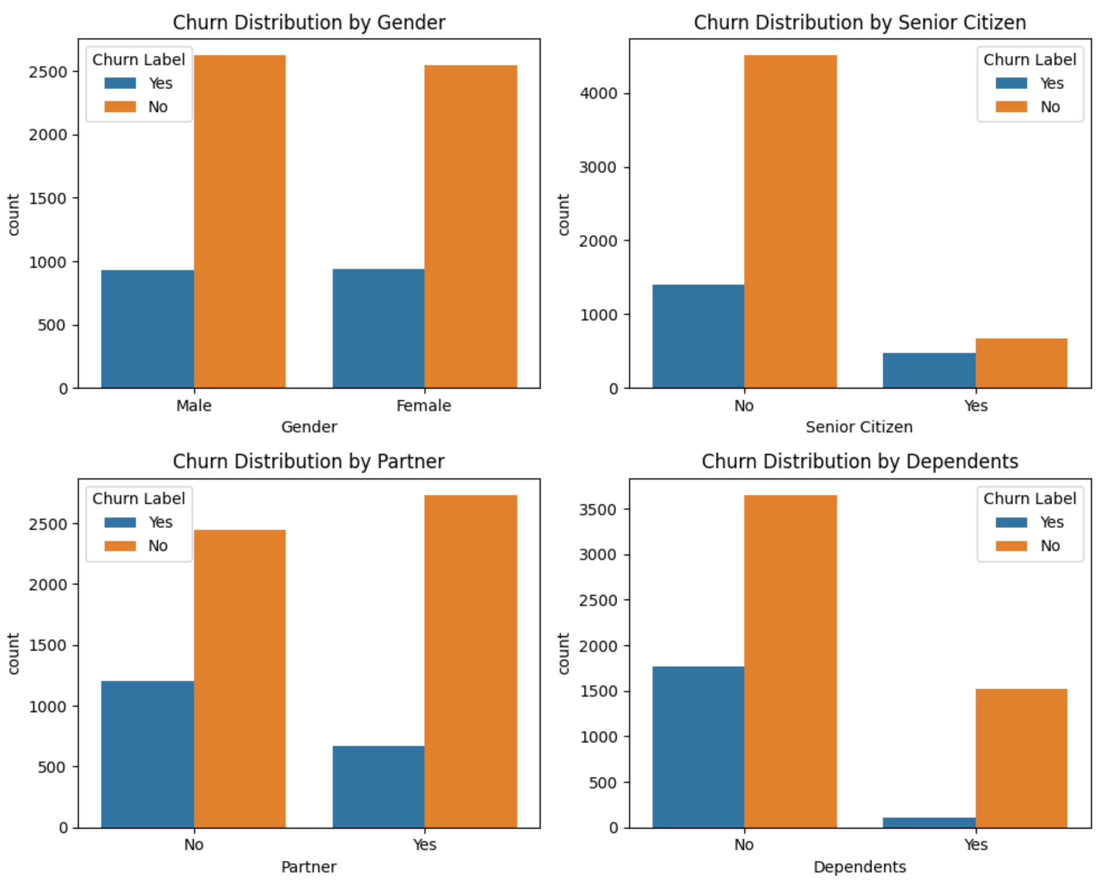


Fig: Explored distribution of churn across various demographic categories.

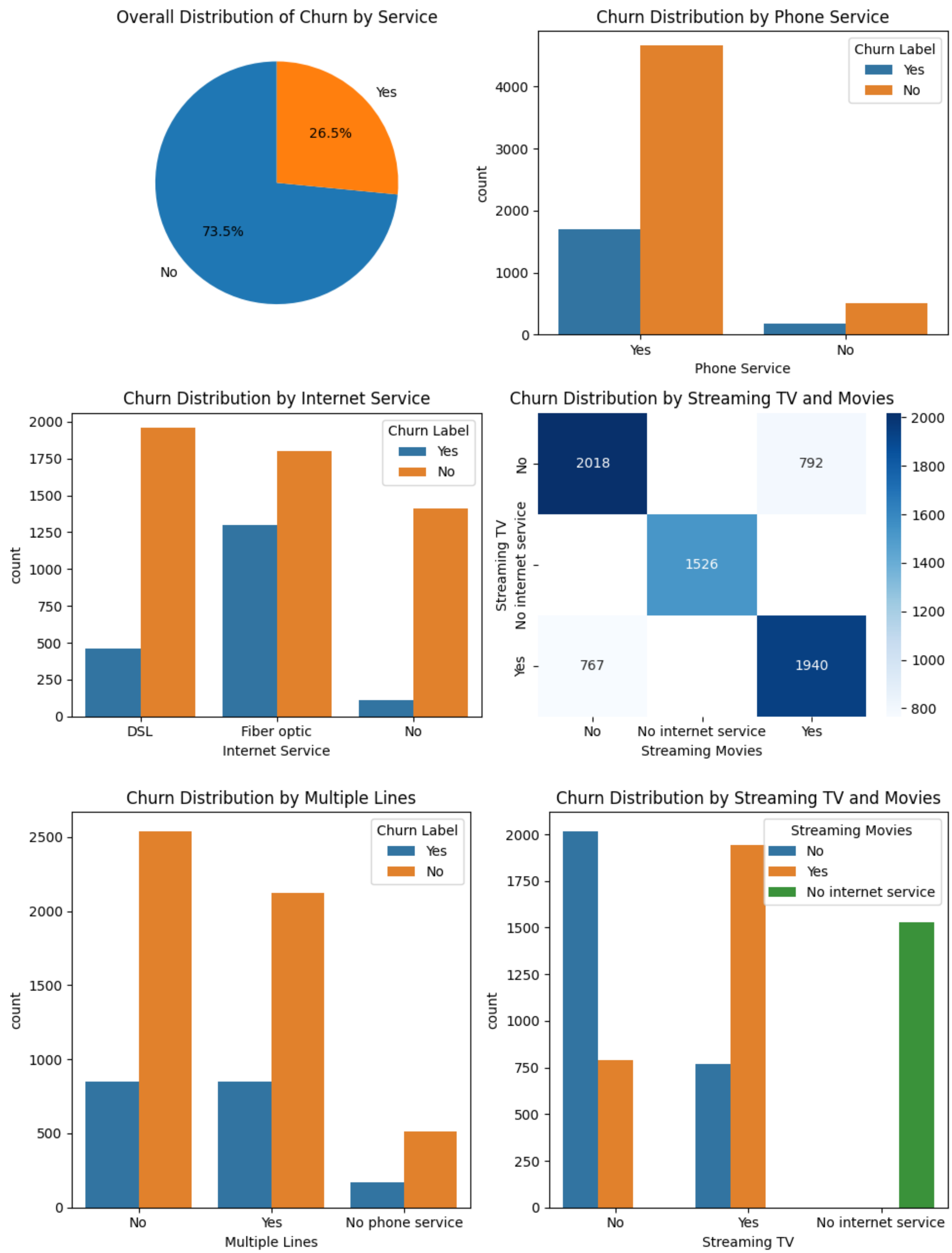


Fig: Explored distribution of churn across various service-related categories.

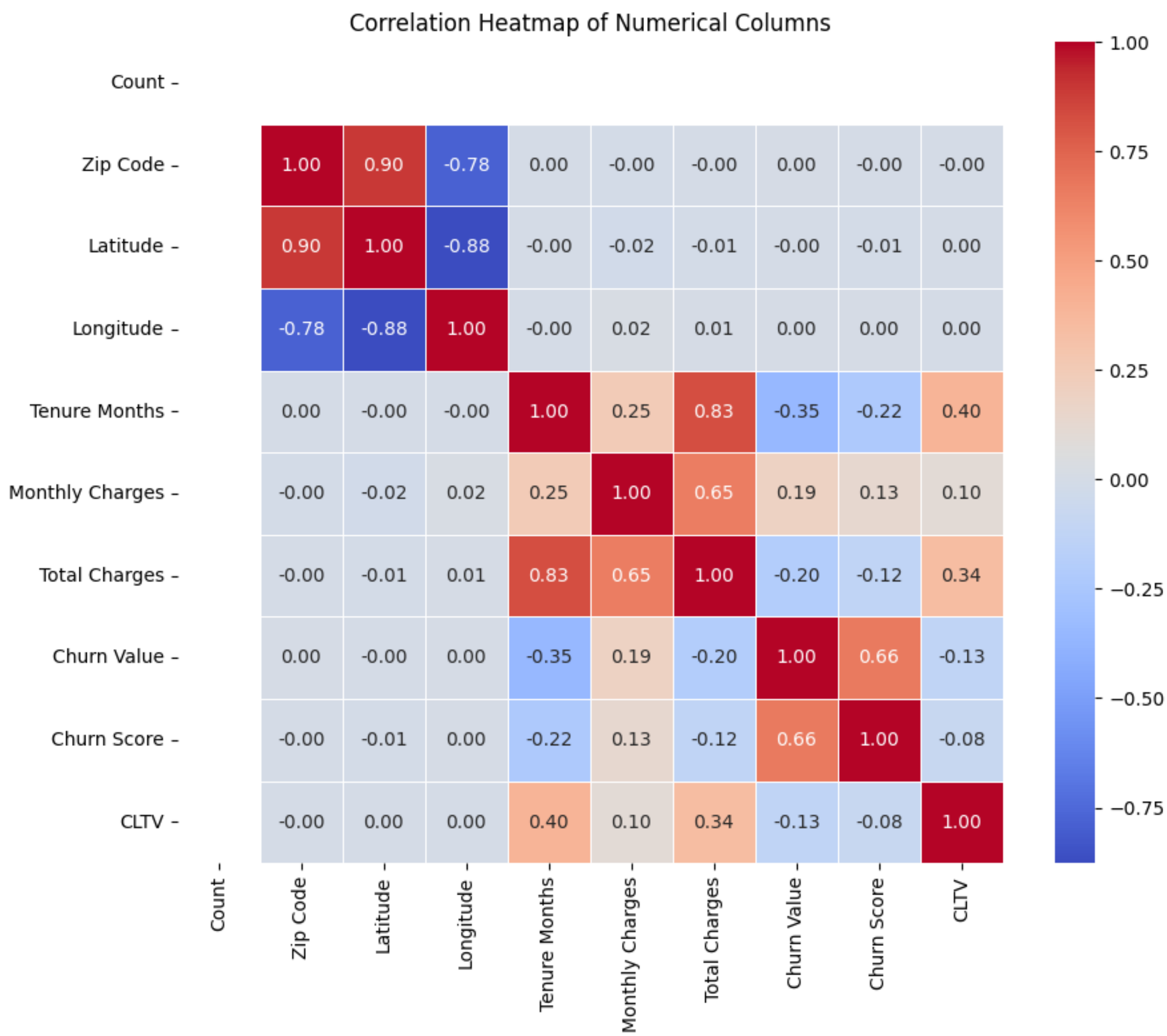


Fig: Correlation Heatmap of Numerical Columns

3. Model Development and Evaluation

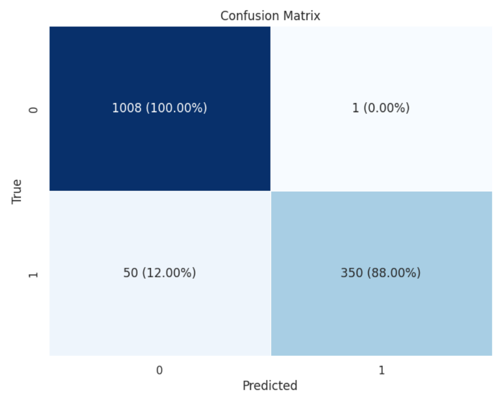
The churn prediction models, including Logistic Regression, Naive Bayes, Random Forest Classifier, and Support Vector Machine (SVM), consistently demonstrate outstanding performance. Accuracy ranges from 96.38% to 99.93%, showcasing the models' ability to effectively predict customer churn. Precision and recall values are well-balanced, affirming the models' reliability in identifying both churn and non-churn cases. These robust results emphasize the models' suitability for deployment in a production environment, enabling businesses to proactively address potential customer churn with high confidence.

3.1. Result:

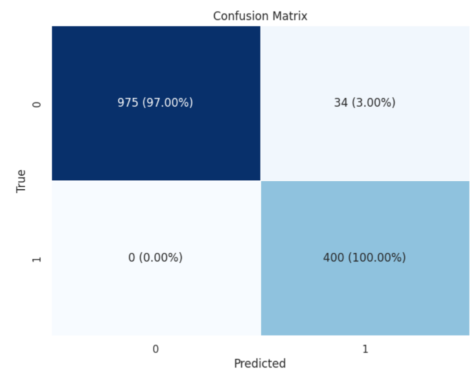
Algorithm	Accuracy	Precision	Recall
Logistic Regression	96.38%	95.00%	88.00%
GaussianNB	97.59%	92.00%	100.00%
Random Forest	99.72% (ROC AUC: 99.48%)	99.00%	100.00%
SVM	99.93%	100.00%	100.00%

Confusion Matrix:

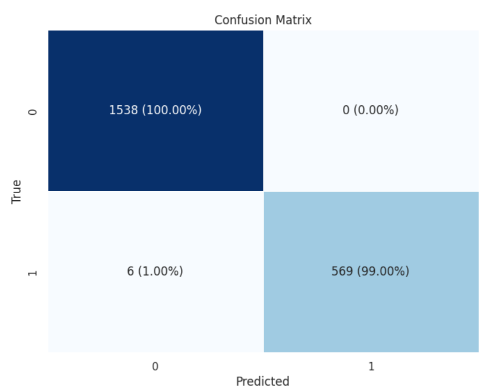
1. Logistic Regression



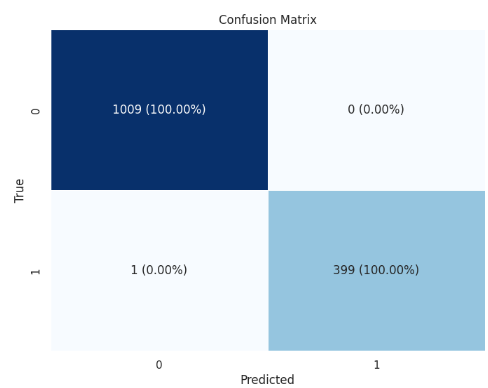
2.GaussianNB



3. Random Forest



4.SVM



3.2.Result Analysis:

1. Random Forest and SVM Performance:

   The Random Forest and Support Vector Machine (SVM) models exhibit outstanding accuracy, precision, and recall in predicting customer churn.

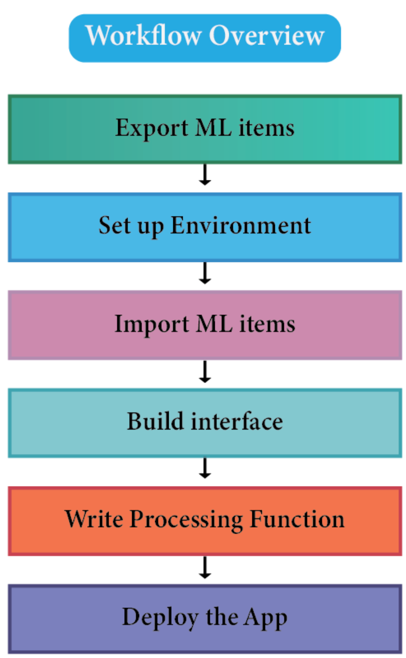
2. Feature Importance:

   A detailed feature importance analysis has been conducted to identify the key predictors contributing to churn in the dataset.

3. Next Steps:

- Consider Deploying SVM for Superior Performance: Given its exceptional performance, particularly in accuracy, precision, and recall, the SVM model is recommended for deployment in predicting customer churn.
- Evaluate Models on Additional Datasets for Generalization: To ensure the models' generalization and robustness, it is advisable to evaluate them on additional datasets. This step will provide insights into how well the models perform across different data distributions and scenarios, enhancing their applicability in real-world situations.

4. Deployment Plan



Process Description: The workflow seamlessly guides the transition from machine learning development to a functional application. Each step contributes to the overall success of the deployment:

1. Export Machine Learning Items from Notebook: Identify and extract relevant machine learning artifacts, such as models and preprocessing steps, from the notebook.
2. Import Machine Learning Items into the App Script: Integrate the exported ML items into the application script, ensuring a cohesive connection between the developed models and the app's functionality.
3. Build an Interface: Create a visually appealing and user-friendly interface, providing a platform for users to input data and receive model predictions or outputs.
4. Write a Function to Process Inputs: Develop a processing function within the application script to handle user inputs. This function should seamlessly connect with the ML model, incorporating any required preprocessing.
5. Pass Values Through the Interface: Enable the interface to capture user inputs and pass them to the backend for further processing.
6. Recover Values in the Backend: Retrieve user inputs in the backend, preparing them for processing before model prediction.
7. Apply Necessary Processing: Implement the required preprocessing steps to ensure that user inputs are in the correct format for model prediction.
8. Submit Processed Values to the ML Model: Feed the processed user inputs into the machine learning model to generate predictions.
9. Process Predictions and Display on Interface: Retrieve and process the predictions obtained from the ML model, displaying the results on the user interface for user consumption.

5. Conclusion

5.1 Achievements:

In the course of this project, we successfully developed and evaluated multiple machine learning models for predicting customer churn. The achieved high accuracy across these models demonstrates their effectiveness in capturing patterns and making reliable predictions regarding customer behavior.

5.2 Future Steps:

Moving forward, a key aspect of maintaining the model's relevance involves continuous monitoring and regular updates. This proactive approach ensures that the model remains aligned with evolving trends and customer behaviors. Additionally, there is potential for enhancing prediction capabilities by incorporating additional features or refining existing ones. By staying attentive to model performance and embracing iterative improvements, we can ensure the sustained success of the churn prediction system.