Customer Churn Prediction Project Report

1. Introduction

This report details the end-to-end process of developing a machine learning model to predict customer churn for a telecommunications company. The primary business objective is to proactively identify customers who are likely to cancel their service, enabling the company to take targeted retention actions.

The project utilizes the classic "Telco Customer Churn" dataset. The modeling approach involves building a sophisticated **stacked ensemble model**, which combines the predictive power of several classic algorithms. The development process is iterative, starting with a baseline model and progressively enhancing its performance through techniques such as class imbalance handling (SMOTE), feature engineering, and hyperparameter tuning (GridSearchCV). The final model's predictions are interpreted using the SHAP (SHapley Additive exPlanations) library to provide actionable, data-driven insights.

---

2. Data Loading and Preprocessing

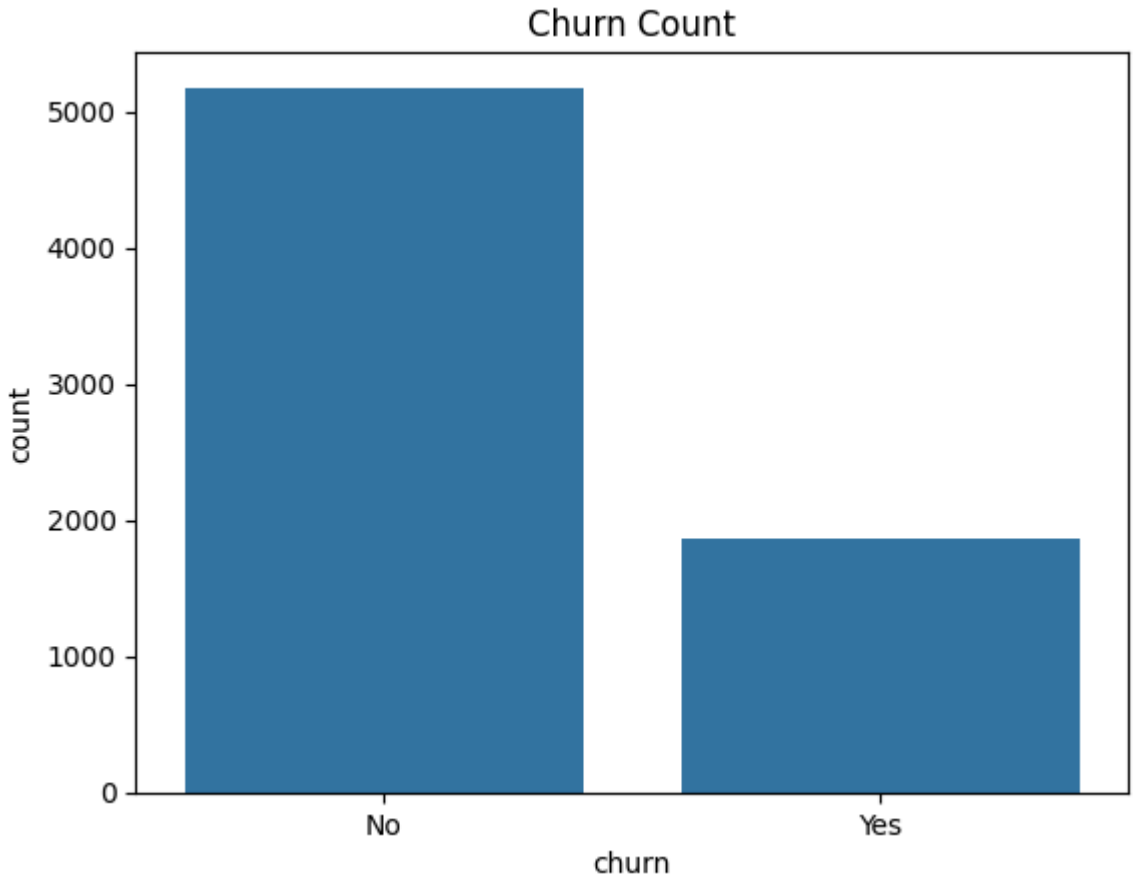
2.1. Data Loading and Initial Cleaning The project began by loading the `WA_Fn-UseC_-Telco-Customer-Churn.csv` dataset using the pandas library. The initial dataset consisted of 7,043 rows and 21 columns.

An initial inspection revealed that the `TotalCharges` column, while expected to be numerical, was registered as an `object` data type. This was due to the presence of empty strings for new customers who had not yet incurred any total charges. This issue was resolved by:

1. Converting the `TotalCharges` column to a numeric type, coercing empty strings into `NaN` (Not a Number) values.
2. Filling these 11 `NaN` values with `0`, as it is logical for a new customer to have zero total charges.

2.2. Exploratory Data Analysis (EDA) EDA was performed to understand the characteristics of the data and identify potential drivers of churn.

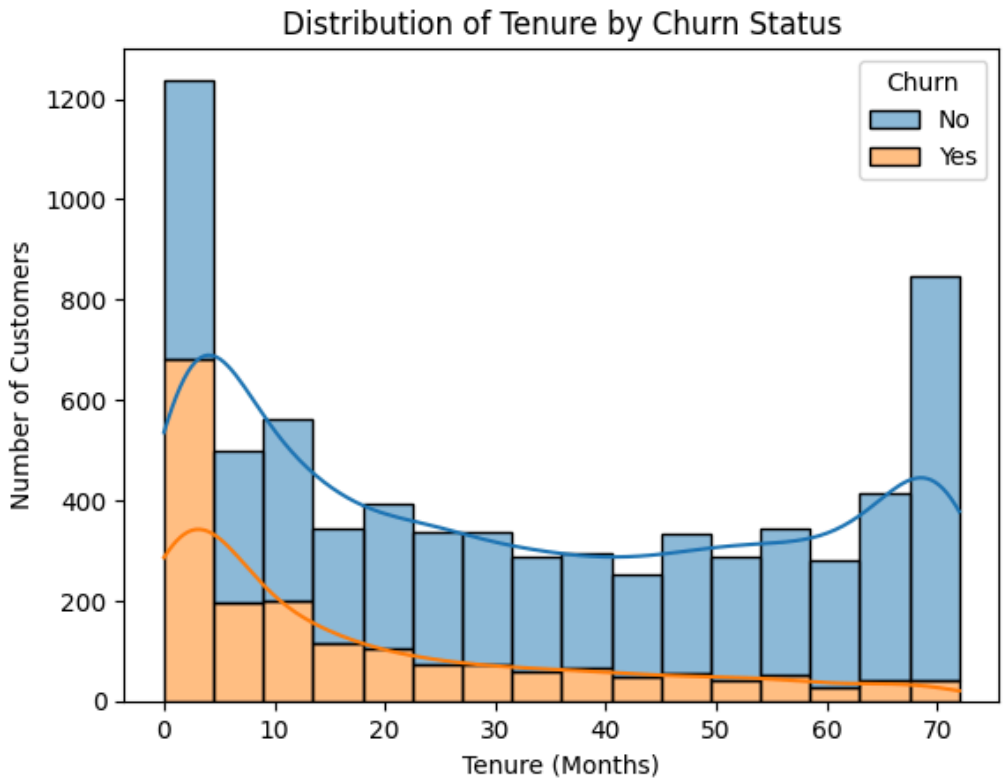
The target variable, `Churn`, was found to be imbalanced, with significantly more customers not churning than churning. This imbalance is a critical consideration for model training and evaluation.



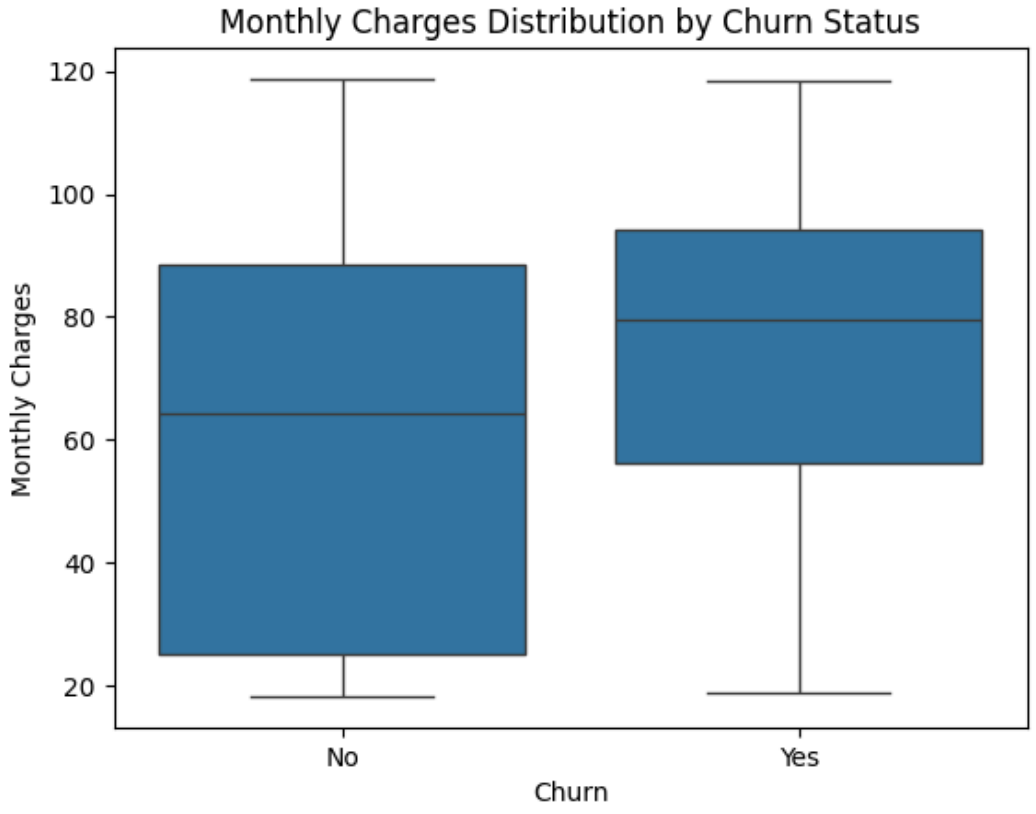
(This plot shows the "No" vs. "Yes" counts for the 'Churn' column.)

Visual analysis of both categorical and numerical features revealed several strong indicators of churn:

- **Contract Type:** Customers with month-to-month contracts have a dramatically higher churn rate compared to those on one or two-year contracts.
- **Tenure:** Newer customers (with low tenure) are far more likely to churn. Long-term customers show much higher loyalty.
- **Internet Service:** Customers with Fiber optic internet service have a notably higher churn rate.
- **Monthly Charges:** Customers with higher monthly charges are more likely to leave.



(This is the histogram plot showing tenure distribution for 'Churn' vs. 'No Churn'.)



(This is the box plot showing the distribution of monthly charges for 'Churn' vs. 'No Churn'.)

---

3. Modeling and Iterative Improvement

A stacked ensemble model was chosen for its high predictive accuracy. The architecture consists of:

- **Level 0 (Base Models):** Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest.
- **Level 1 (Meta-Model):** XGBoost, which learns to optimally combine the predictions from the base models.

3.1. Baseline Model Evaluation The initial model was trained on the preprocessed data with `class_weight='balanced'` to partially address the class imbalance. While the overall accuracy was decent at 78%, the performance on the minority class ('Churn') was poor, with a **recall of only 48%**. This means the baseline model failed to identify more than half of the customers who were actually churning.

Classification Report (Baseline Model):

	precision	recall	f1-score	support
0	0.82	0.89	0.86	1035
1	0.61	0.48	0.54	374
accuracy			0.78	1409

3.2. Improvement 1: Handling Class Imbalance with SMOTE To address the low recall, the **Synthetic Minority Over-sampling Technique (SMOTE)** was applied to the training data. SMOTE balances the class distribution by creating synthetic examples of the minority class. This allows the model to learn the characteristics of churning customers more effectively. After re-training on the SMOTE-balanced data, the **recall for the 'Churn' class significantly improved from 48% to 55%**.

Classification Report (After SMOTE):

	precision	recall	f1-score	support
0	0.84	0.83	0.83	1035
1	0.54	0.55	0.54	374
accuracy			0.76	1409

3.3. Improvement 2: Feature Engineering To provide the model with more predictive signals, three new features were engineered:

1. **ServicesUsed:** A count of the optional services each customer subscribes to.
2. **TenureInYears:** The customer's tenure converted to years.
3. **MonthToTotalRatio:** The ratio of `MonthlyCharges` to `TotalCharges`, which can highlight new, high-paying customers.

The model was re-trained using these new features (along with SMOTE). This step maintained the high recall while slightly improving precision.

3.4. Final Model: Hyperparameter Tuning with GridSearchCV The final step in optimization was to tune the hyperparameters of the XGBoost meta-model using `GridSearchCV`. The search was optimized for the **'recall'** metric to ensure the final model was best at identifying at-risk customers. This process yielded the best-performing model of the project.

Final Model Classification Report:

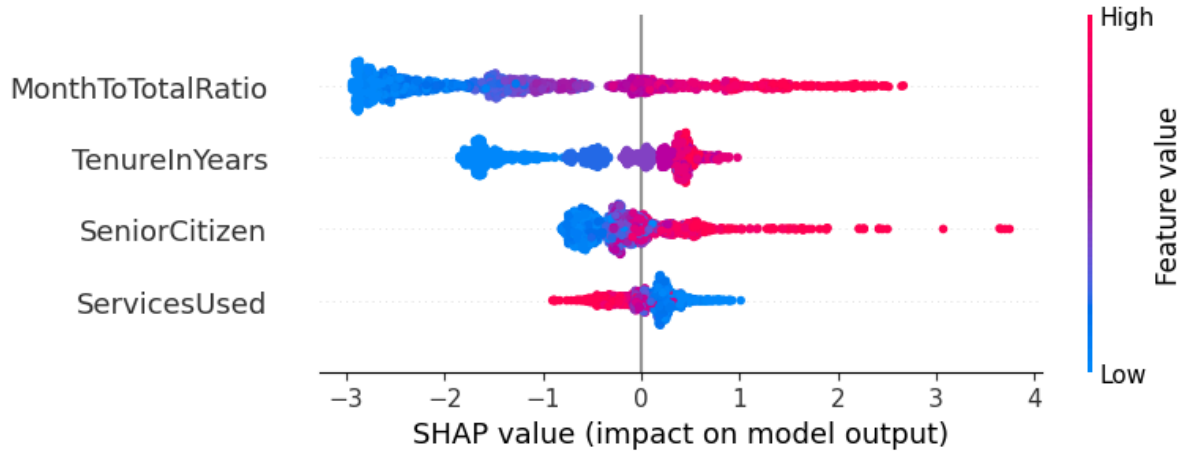
	precision	recall	f1-score	support
0	0.84	0.83	0.84	1035
1	0.55	0.57	0.56	374
accuracy			0.76	1409

The final model achieved a balanced performance, with a **churn recall of 57%** and a **churn precision of 55%**, representing a significant improvement over the baseline.

---

4. Model Interpretability with SHAP

To understand the key drivers behind the final model's predictions, a SHAP summary plot was generated. The analysis confirmed the findings from the EDA and highlighted the power of the newly engineered features.



(This is the bee swarm plot from the final, tuned model.)

Key Insights from SHAP:

- The **`MonthToTotalRatio`** was the single most impactful feature, indicating that new, high-paying customers are the highest churn risk.
- **`TenureInYears`** was also highly influential, confirming that long-term customers are much more loyal.

- The number of **`ServicesUsed`** was a key factor, showing that customers more deeply integrated into the service ecosystem are less likely to churn.

---

5. Model Saving and Experiment Tracking

For deployment and reproducibility, the following steps were taken:

- The final trained `best_model` object, along with the `StandardScaler` and training column lists, were saved to disk using `joblib`.
- The final experiment's parameters and performance metrics were logged using **MLflow**. This ensures that the model development process is well-documented and the best model can be easily retrieved for future use.

---

6. Conclusion

This project successfully developed a robust and interpretable stacked ensemble model for predicting customer churn. Through a systematic process of EDA, iterative modeling, feature engineering, and hyperparameter tuning, the model's ability to identify churning customers (recall) was improved from an initial 48% to a final 57%. The SHAP analysis provided clear, actionable insights into the main factors driving churn, with the engineered features proving to be highly predictive. The final model and associated artifacts have been saved and are ready for deployment in a production environment.