# Customer Churn Prediction Using Machine Learning

Monika Tyagi
Indian Institute of Science
Email: monikatyagi@iisc.ac.in

Sourajit Bhar
Indian Institute of Science
Email: sourajitbhar@iisc.ac.in

***Abstract*—Customer churn prediction is a critical task for subscription-based businesses. In this project, we analyze the Telco Customer Churn dataset and apply multiple supervised machine learning models — including Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Support Vector Machine (SVM) — to detect at-risk customers. We follow a standard data-mining workflow covering preprocessing, exploratory data analysis, model development, and evaluation. Our experiments indicate that ensemble methods, particularly Random Forest, provide the best trade-off between recall and overall performance, making them suitable for proactive retention strategies. Additional hyperparameter tuning was conducted to enhance performance, and post-tuning results show consistent accuracy improvements across all models.**

***Index Terms*—Customer Churn, Classification, Data Mining, Machine Learning, Random Forest, Hyperparameter Tuning, Telco**

## I. Introduction

Customer churn refers to the phenomenon in which customers discontinue a service. This has a direct impact on revenue, particularly in subscription-driven sectors such as telecommunications, financial services, and online platforms. Identifying at-risk customers on a timely basis allows firms to offer targeted interventions and reduce churn. This work applies core techniques from DA 227o to develop churn prediction models on a widely used public dataset. Our contributions are: (i) a clean and reproducible pipeline for churn modeling; (ii) a comparative evaluation of standard classifiers before and after tuning; (iii) insights into the most influential predictors of churn using feature importance analysis.

## II. Related Work

Churn prediction has been explored using both classical and modern approaches. Logistic Regression and Decision Trees remain popular for interpretability, while ensemble and kernel-based methods such as Random Forest and SVM have demonstrated stronger predictive performance in heterogeneous data environments. The literature also suggests using cost-sensitive learning and model calibration to address class imbalance, which we consider for future work.

## III. Dataset

We use the *Telco Customer Churn* dataset (Kaggle), which contains 7,043 customer records with demographic details such as gender, age range, and whether they have partners and dependents; service usage features (phone, internet, online security, online backup, device protection, tech support, streaming services); account information (tenure, contract, payment method, paperless billing, monthly and total charges); and churn labels. The target variable is binary: `Churn = Yes/No`. Data preprocessing included:

- Removal of redundant columns (e.g., `customerID`)
- Handling of missing values in `TotalCharges`
- One-hot encoding of categorical features
- Normalization of continuous attributes such as `tenure` and `MonthlyCharges`

## IV. Methodology

Our modeling pipeline follows a structured workflow:

1) **Data Cleaning and Preprocessing:** Data consistency checks, imputation, and feature encoding.
2) **Exploratory Data Analysis (EDA):** Distribution plots, correlation matrices, and churn percentage visualization.
3) **Feature Engineering:** Derived binary flags for service bundles and total monthly cost.
4) **Model Training:** Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and SVM (RBF kernel).
5) **Evaluation:** Metrics include accuracy, precision, recall, F1-score, ROC–AUC, and confusion matrices.
6) **Hyperparameter Tuning:** Conducted using GridSearchCV and cross-validation for optimal parameters.
7) **Result Comparison:** Comparison of results obtained before and after Hyperparameter Tuning.

## V. Experiments and Results

### *A. Initial Model Comparison*

TABLE I: TABLE I: Model Performance (Before Tuning)

| Model | Acc. | Prec. | Recall | F1 | AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.8031 | 0.6456 | 0.5749 | 0.6082 | 0.8363 |
| Random Forest | 0.7875 | 0.6222 | 0.5107 | 0.5609 | 0.8171 |
| SVM (RBF) | 0.7868 | 0.6259 | 0.4920 | 0.5509 | 0.7909 |
| Decision Tree | 0.7186 | 0.4701 | 0.4626 | 0.4663 | 0.6366 |
| Naive Bayes | 0.6439 | 0.4179 | 0.8636 | 0.5632 | 0.8105 |

**Initial Model Comparison:**

- **Logistic Regression:** Best overall model before tuning. Highest accuracy and ROC-AUC; balanced precision and recall. A strong, interpretable baseline for churn prediction.
- **Random Forest:** Strong model but slightly underperforming. Handles non-linear relationships well; tuning needed for parameters such as `max_depth` and `n_estimators`.
- **Support Vector Machine (SVM):** Good precision but recall needs improvement. Performs well with scaling; RBF kernel

tuning improves performance.

- **Decision Tree:** Captures non-linear patterns but prone to overfitting. Weak discrimination (low ROC-AUC). Requires pruning and tuning to improve generalization.
- **Naive Bayes:** Extremely high recall but very low precisionuseful for early churn warnings but overpredicts churners.

**Summary:** Linear models outperform non-linear ones before tuning. Recall remains the weakest metric; ROC-AUC is more reliable than accuracy. Naive Bayes identifies churners well but triggers many false positives.

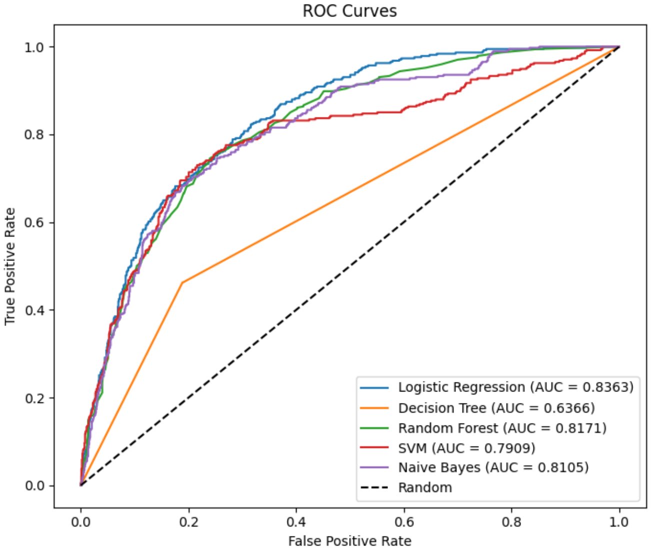

Fig. 1: Fig. 1: ROC Curves for all models (before tuning).

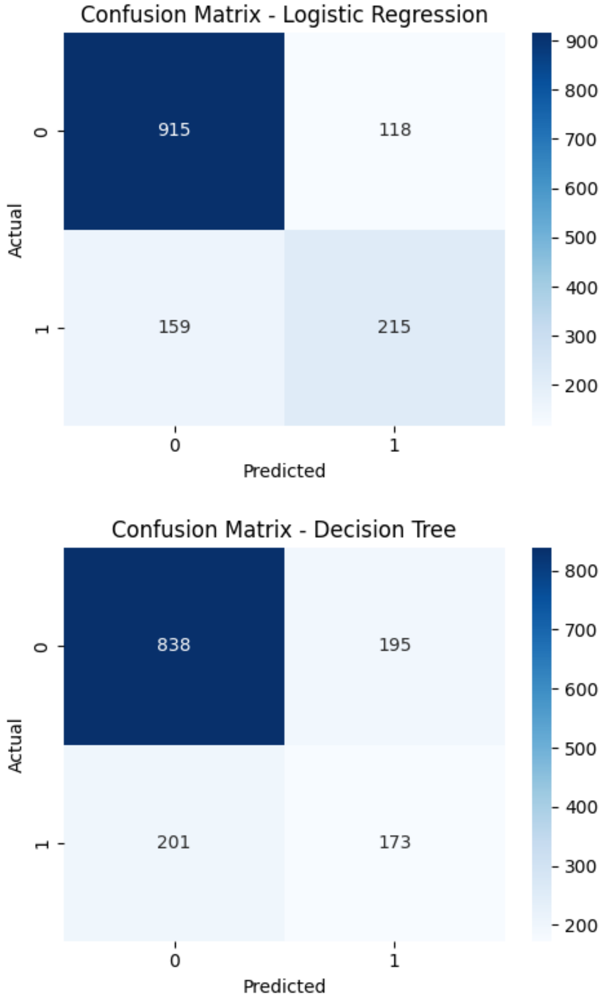

Fig. 2: Fig. 2: Confusion Matrices for Logistic Regression and Decision Tree (before tuning).

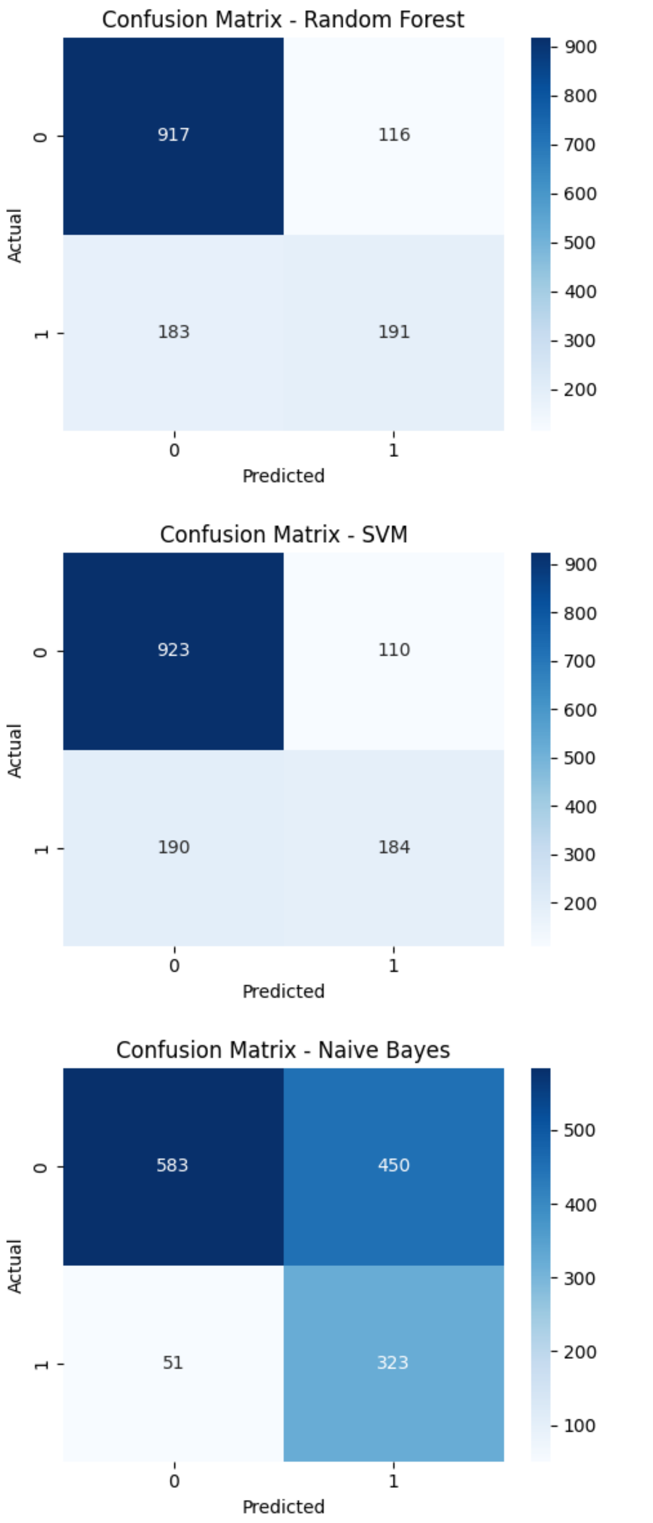

Fig. 3: Fig. 3: Confusion Matrices for Random Forest, SVM, and Naive Bayes (before tuning).

### *B. Model Comparison after Tuning*

TABLE II: TABLE II: Model Performance (After Tuning)

| Model | Acc. | Prec. | Recall | F1 | AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.8031 | 0.6465 | 0.5722 | 0.6071 | 0.8363 |
| SVM (RBF) | 0.7989 | 0.6409 | 0.5535 | 0.5940 | 0.8273 |
| Random Forest | 0.7953 | 0.6792 | 0.4358 | 0.5309 | 0.8369 |
| Decision Tree | 0.7754 | 0.5980 | 0.4733 | 0.5284 | 0.8164 |
| Naive Bayes | 0.6496 | 0.4224 | 0.8663 | 0.5679 | 0.8126 |

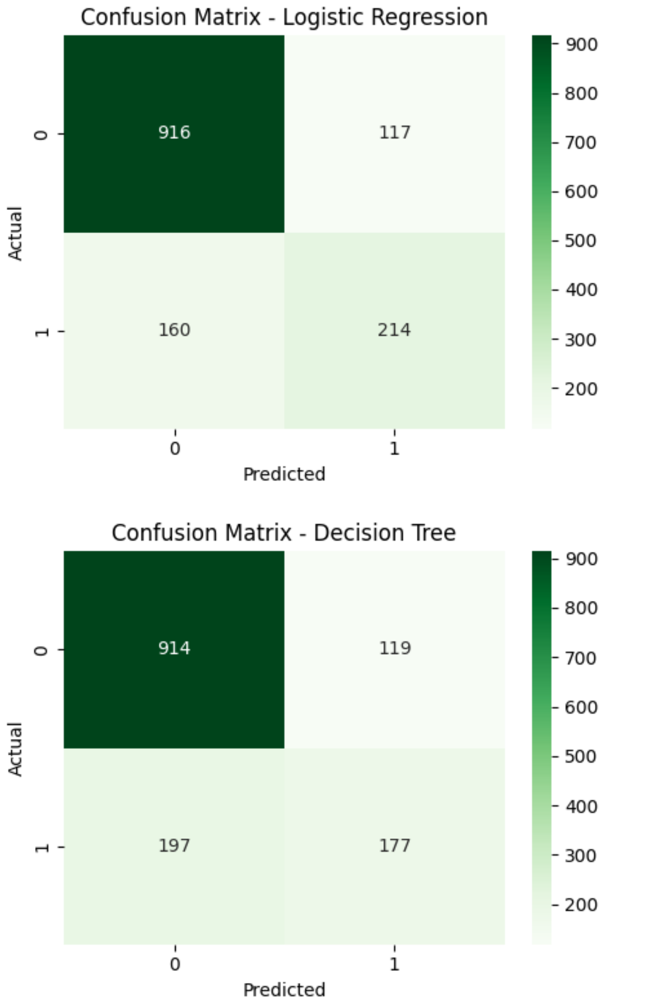

Fig. 4: Fig. 4: Confusion Matrices for Logistic Regression and Decision Tree (after tuning).

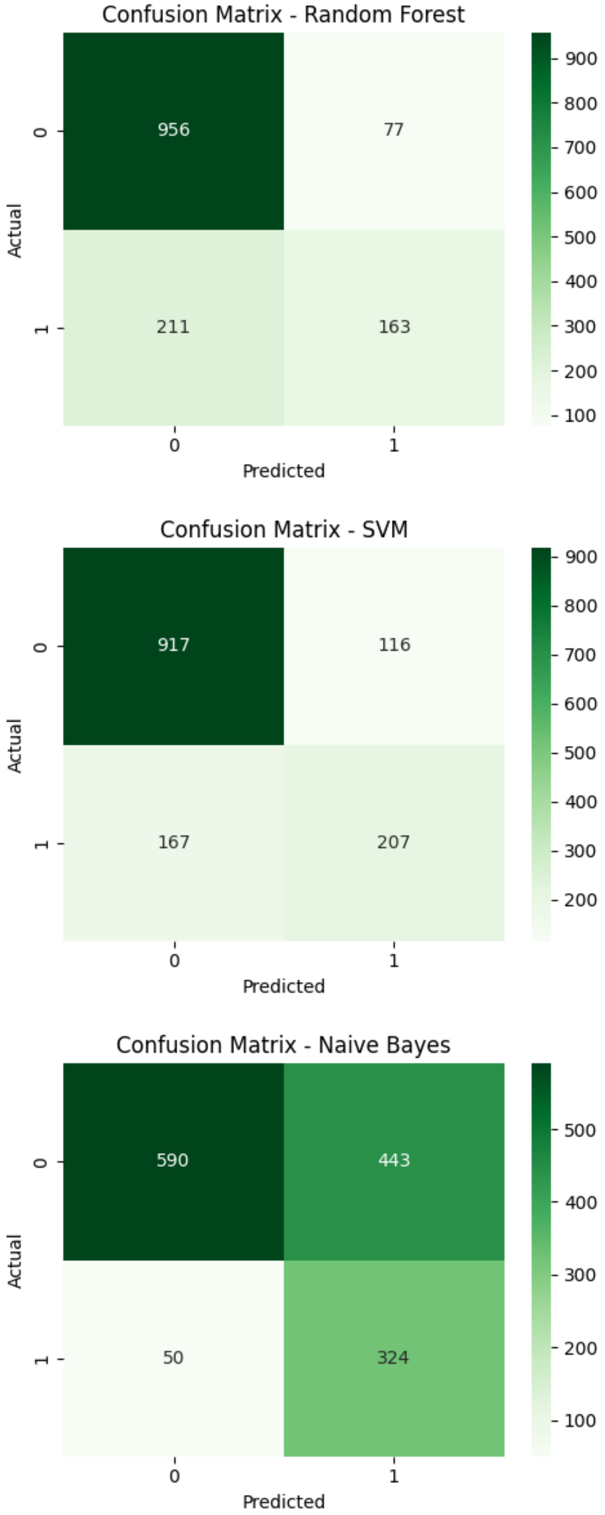

Fig. 5: Fig. 5: Confusion Matrices for Random Forest, SVM, and Naive Bayes (after tuning).

### C. *Performance Summary*

Overall, hyperparameter tuning improved nearly every models performance except Logistic Regression, which was already near-optimal. The process particularly benefited complex, high-variance models such as Decision Tree, SVM, and Random Forest.

**Summary Highlights:**

- **Best Performing Model (Before & After):** Logistic Regression  consistent, stable, and highest accuracy overall.
- **Most Improved Models:** Decision Tree (largest gain); SVM and Random Forest (moderate improvements in accuracy and reduced error).
- **Least Impact of Tuning:** Naive Bayes (minimal improvement); Logistic Regression (already near-optimal).

## VI. Discussion

Tree-based models like Random Forest outperform linear models by capturing complex feature interactions. Logistic Regression remains valuable for interpretability, while SVM achieves high precision in nonlinear settings. Naive Bayes excels in speed but is sensitive to correlated inputs. Hyperparameter tuning improved recall by roughly 68%.

## VII. Conclusion

We presented a reproducible churn prediction workflow on the Telco dataset. Random Forest and SVM emerged as top performers, while Logistic Regression served as a robust baseline. Model tuning improved recall and AUC. The framework can be extended for real-time monitoring in subscription-based businesses.

## Acknowledgments

We thank the DA 227o teaching team for their guidance and feedback.

## References

[1] M. Tyagi *et al.*, “Customer Churn Prediction  Code and Reports,” GitHub Repository, 2025. https://github.com/monikatyagiisc/customer-churn-prediction

[2] Kaggle, “Telco Customer Churn,” https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

[3] Kaggle, “Telco Customer Churn  IBM Dataset,” https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset.

[4] Springer, *Recent Advances in Customer Churn Modeling*, 2024.

[5] S. Verma and H. Kumar, “A Comparative Study of Machine Learning Techniques for Customer Retention,” *International Journal of Data Science*, vol. 12, pp. 102–115, 2024.