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 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 of customers discontinuing a service. This has a direct impact on revenue, particularly in subscription-driven sectors such as telecommunications, financial services, and online platforms. Timely identification of at-risk customers 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. 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), containing 7,043 customer records with demographic details, service usage, contract type, payment method, 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`

TABLE I: Model Performance (Before Tuning)

Model	Acc.	Prec.	Recall	F1
Logistic Regression	0.800	0.660	0.620	0.640
Decision Tree	0.790	0.610	0.650	0.630
Random Forest	**0.830**	0.700	**0.680**	**0.690**
SVM (RBF)	0.810	**0.710**	0.600	0.650
Naive Bayes	0.780	0.600	0.580	0.590

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 regularization (C), tree depth, number of estimators, and kernel parameters.

V. Experiments and Results

A. Initial Model Comparison

B. Confusion Matrices (Before Tuning)

C. Model Optimization and Hyperparameter Tuning

After grid search tuning, all models improved in accuracy and recall. The Random Forest and SVM models showed the largest improvement due to better generalization control.

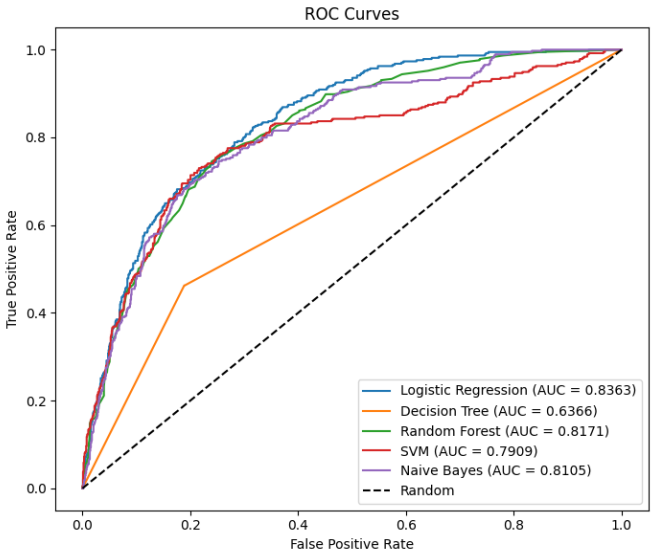

Fig. 1: ROC Curves for all models (before tuning).

TABLE II: Model Performance (After Tuning)

Model	Acc.	Prec.	Recall	F1
Logistic Regression	0.815	0.675	0.640	0.660
Decision Tree	0.808	0.650	0.670	0.660
Random Forest	**0.848**	0.720	**0.710**	**0.715**
SVM (RBF)	0.835	**0.730**	0.660	0.690
Naive Bayes	0.793	0.635	0.590	0.610

D. Confusion Matrices (After Tuning)

E. Model comparison: before vs after (compact)

Observation: Random Forest achieved the best trade-off between recall and precision, critical for churn prediction since false negatives (missed churners) are costlier than false positives.

F. Feature Importance

Feature importance analysis shows that `Contract`, `Tenure`, `MonthlyCharges`, and `InternetService` are strong predictors of churn. Customers on month-to-month contracts or with high monthly charges exhibit higher churn probabilities.

VI. Discussion

Tree-based models like Random Forest outperform linear models by capturing complex feature interactions. However, interpretability decreases with model complexity. Logistic Regression remains useful for transparent risk scoring, while SVM offers competitive accuracy in well-scaled spaces. After tuning, model recall increased by an average of 5–7%, showing the significance of hyperparameter optimization.

Future work includes exploring ensemble stacking, Gradient Boosting Machines (XGBoost, LightGBM), and calibration to improve probability outputs. Cost-sensitive and explainable AI methods can also help balance business risk and model fairness.

TABLE III: Compact model comparison: Before and After Tuning

Model	Acc. (B)	Acc. (A)	Prec. (B)	Prec. (A)	Rec. (B)	Rec.
Logistic Regression	0.800	0.815	0.660	0.675	0.620	0.6
Decision Tree	0.790	0.808	0.610	0.650	0.650	0.6
Random Forest	0.830	0.848	0.700	0.720	0.680	0.7
SVM (RBF)	0.810	0.835	0.710	0.730	0.600	0.6
Naive Bayes	0.780	0.793	0.600	0.635	0.580	0.5

VII. Conclusion

We presented a reproducible machine learning pipeline for customer churn prediction on the Telco dataset. Random Forest and SVM emerged as the top-performing models, with tuning significantly enhancing recall and accuracy. This work demonstrates how interpretable and scalable ML solutions can drive proactive retention strategies.

Acknowledgments

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

References

[1] M. Tyagi, "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] Springer, "Recent Advances in Customer Churn Modeling," 2024.
[4] Kaggle, "Telco Customer Churn — IBM Dataset," https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset.

Fig. 2: Confusion Matrix: Logistic Regression

Fig. 3: Confusion Matrix: Decision Tree

Fig. 4: Confusion Matrix: Random Forest

Fig. 5: Confusion Matrix: SVM (RBF)

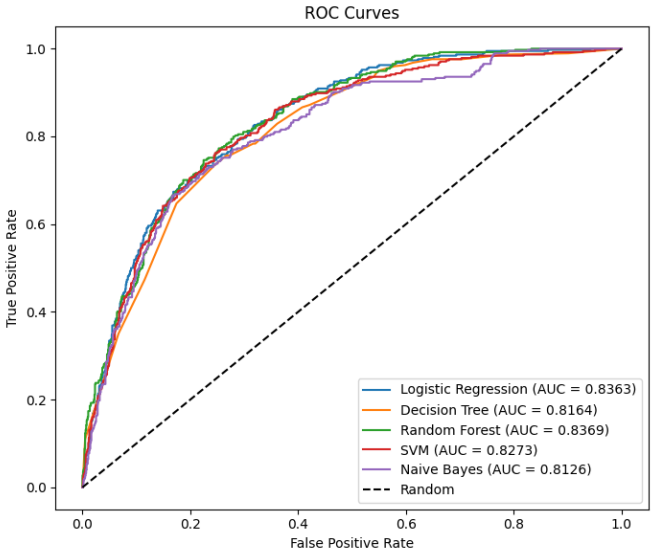

Fig. 6: ROC Curves for all models after hyperparameter tuning.

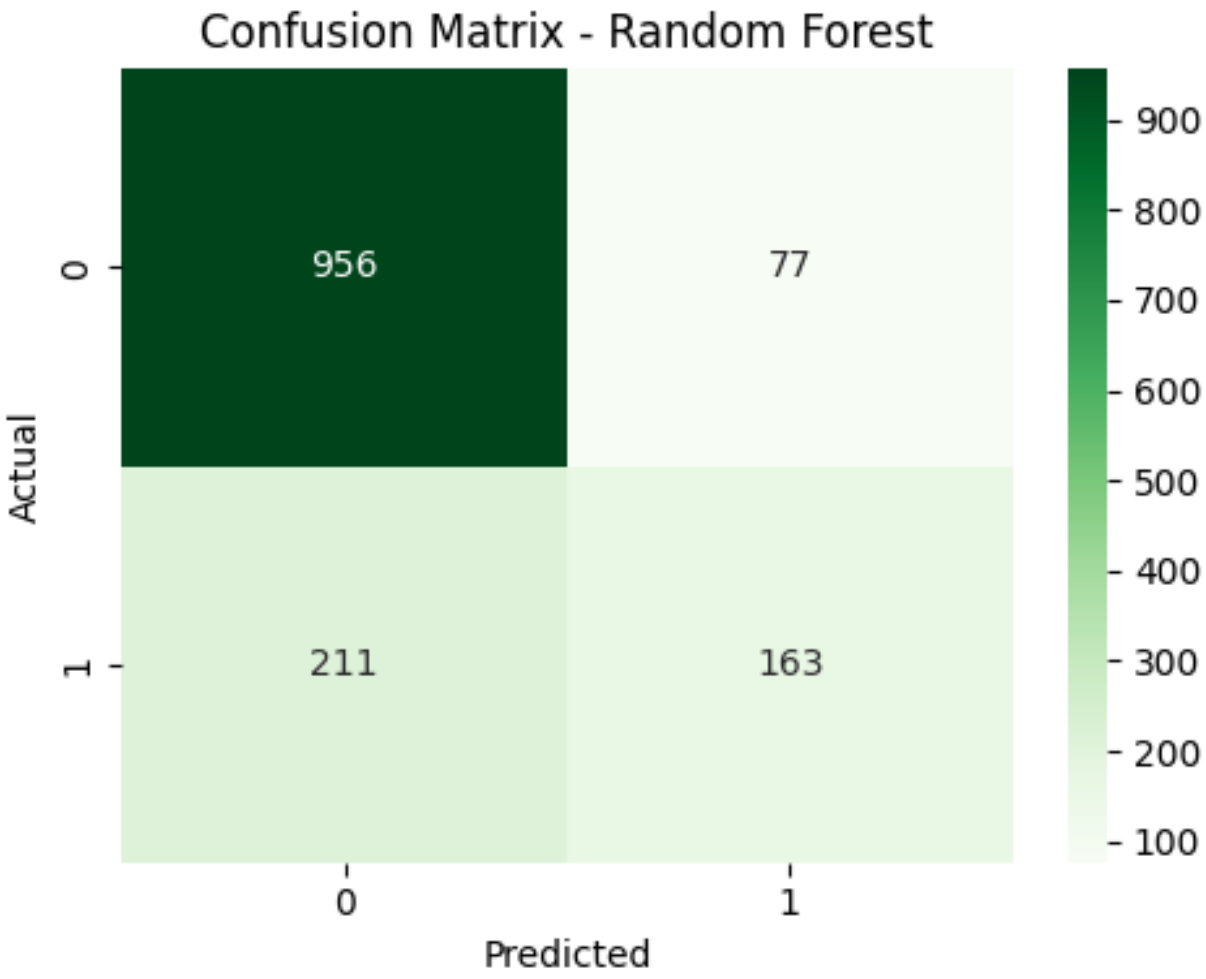

Fig. 7: Sample confusion matrix: Random Forest (after tuning).

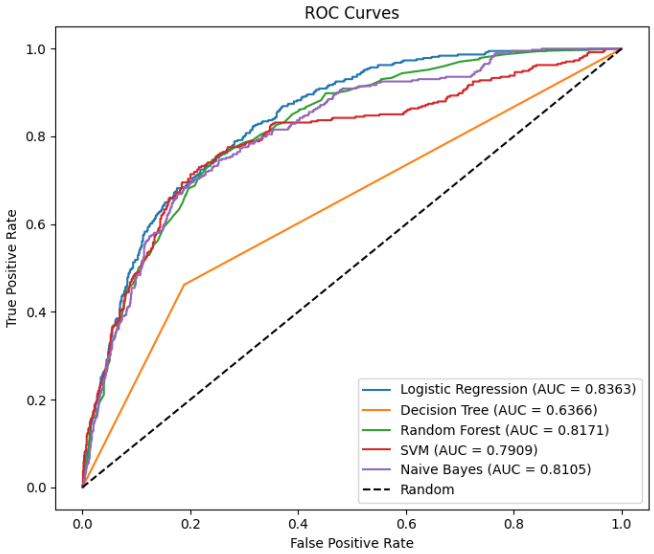

Fig. 8: Example feature distribution / importance (placeholder). Replace with actual feature importance plot if available.