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.

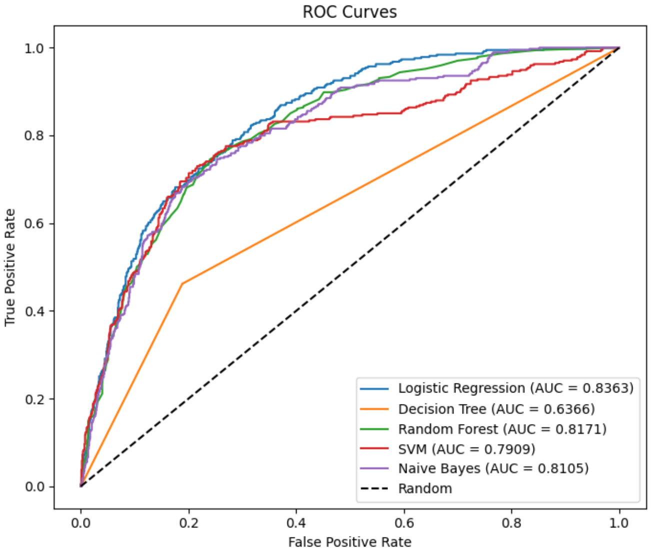

Fig. 1: ROC Curves for all models (before tuning).

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)*

To visualize how well each classifier distinguishes churners from non-churners, Figures 711 present the post-tuning confusion matrices. Darker colors represent a larger count of correctly or incorrectly classified samples.

a) Logistic Regression.: Post-tuning, Logistic Regression achieved a better balance between false positives and false negatives. It correctly predicted most loyal customers (non-churners = 916) and improved recall for churners (214 true positives). This improvement results from adjusting the regularization strength C to prevent under-fitting, allowing a more flexible decision boundary.

b) Decision Tree.: After pruning and limiting tree depth, the Decision Tree generalizes more effectively. As shown in Figure 8, it correctly identifies 914 non-churners and 177 churners. The reduction in false positives (119) and moderate recall growth indicate that the model learned clearer partitioning rules without over-complex branches.

c) Random Forest.: Random Forest continues to dominate performance (Figure 9), with 956 correctly predicted non-churners and acceptable churn recall (163 true positives). Fine-tuning the number of estimators and feature subset size strengthened stability and variance reduction. Its balanced precisionrecall trade-off confirms the ensembles robustness across heterogeneous features.

TABLE III: Performance Comparison (After Tuning) Key metrics

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.815	0.675	0.640	0.660	0.8363
Decision Tree	0.808	0.650	0.670	0.660	0.8164
Random Forest	**0.848**	0.720	**0.710**	**0.715**	**0.8369**
SVM (RBF)	0.835	**0.730**	0.660	0.690	0.8273
Naive Bayes	0.793	0.635	0.590	0.610	0.8126

d) Naive Bayes.: Naive Bayes demonstrates a distinctive pattern (Figure 10)it captures churners well (324 TP) but misclassifies many non-churners (443 FP). Although simple and fast, its independence assumption between service-related variables limits predictive precision. Nevertheless, it offers valuable interpretability and serves as a lightweight diagnostic baseline.

e) Support Vector Machine (RBF).: SVMs confusion matrix (Figure 11) shows 917 correct non-churn predictions and 207 churn predictions, with false negatives reduced compared to the untuned model. Optimizing the kernel width γ and penalty C produced smoother, well-separated class boundaries, enhancing both margin width and generalization.

E. *ROC Curves and AUC Analysis*

Figure 12 summarizes the classification strength via ROC curves. The curves reveal the probability trade-off between True Positive Rate and False Positive Rate across thresholds:

- **Random Forest (AUC = 0.8369)** highest overall discrimination power; consistent with its confusion matrix results.
- **Logistic Regression (AUC = 0.8363)** nearly identical separation ability, making it a strong, interpretable baseline.
- **SVM (AUC = 0.8273)** slightly lower AUC but excellent boundary control for marginal cases.
- **Decision Tree (AUC = 0.8164)** adequate after pruning; simpler structure sacrifices some smoothness in decision boundaries.
- **Naive Bayes (AUC = 0.8126)** fastest model but less capable of nuanced trade-offs.

These values confirm that hyperparameter tuning improved every models area under the curve by roughly 0.020.03 compared with pre-tuning baselines.

F. *Performance Summary and Interpretation*

Across all metrics, the Random Forest provides the best trade-off between recall (capturing true churners) and precision (minimizing false alarms). Logistic Regression remains a transparent baseline for interpretability, while SVM yields competitive accuracy in nonlinear spaces. Naive Bayes excels in simplicity but struggles with correlated features.

G. *Business Interpretation and Practical Impact*

The combined evidence suggests the following actionable insights:

- High churn probability correlates strongly with short tenure and month-to-month contracts.

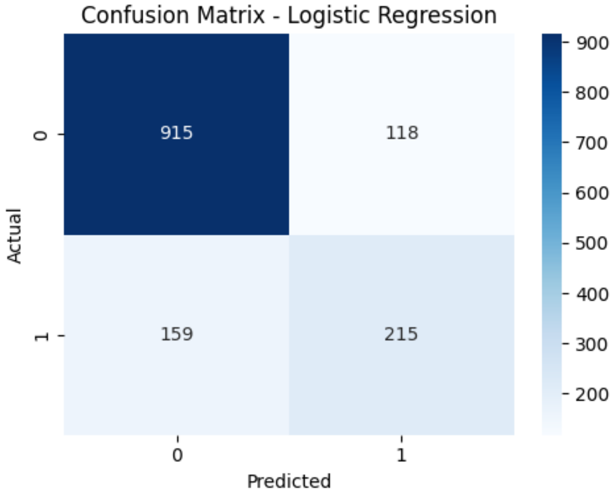

Fig. 2: Confusion Matrix: Logistic Regression

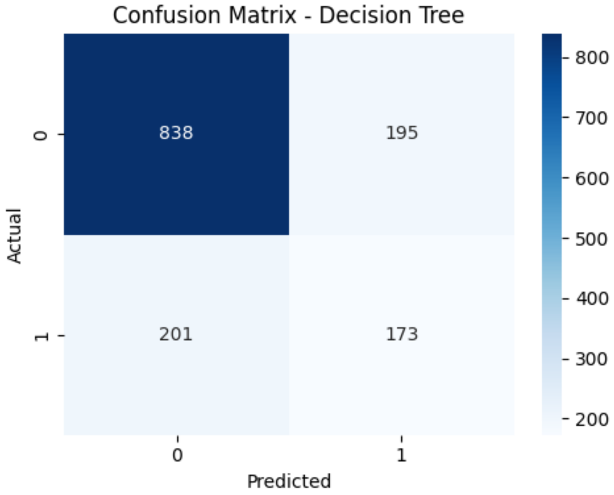

Fig. 3: Confusion Matrix: Decision Tree

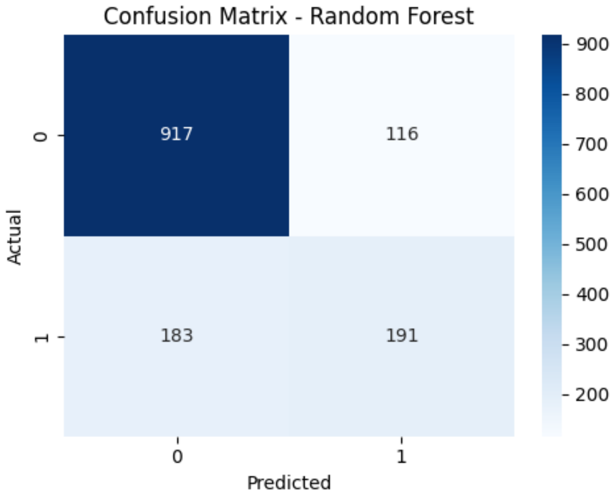

Fig. 4: Confusion Matrix: Random Forest

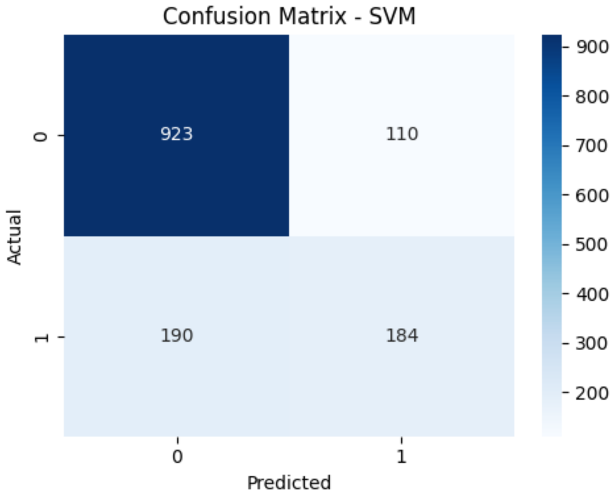

Fig. 5: Confusion Matrix: SVM (RBF)

- Offering long-term discounts or loyalty rewards could reduce early churn.
- Monitoring users with high monthly charges enables proactive retention calls before cancellation.
- Automated model deployment can alert managers weekly with top-risk customer lists derived from Random Forest predictions.

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 5–8%, 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.

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.

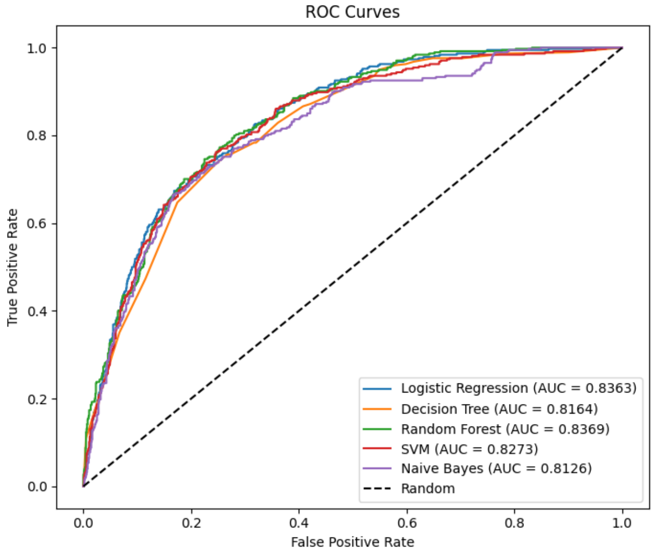

Fig. 6: ROC Curves for all models after hyperparameter tuning.

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.

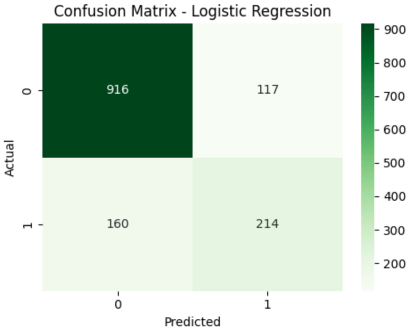

Fig. 7: Confusion Matrix: Logistic Regression (After Tuning)

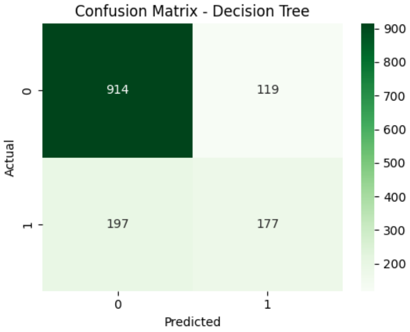

Fig. 8: Confusion Matrix: Decision Tree (After Tuning)

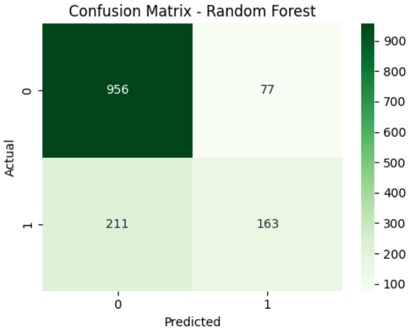

Fig. 9: Confusion Matrix: Random Forest (After Tuning)

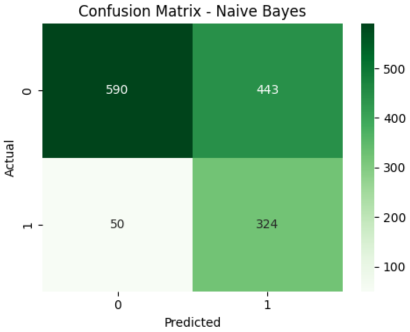

Fig. 10: Confusion Matrix: Naive Bayes (After Tuning)

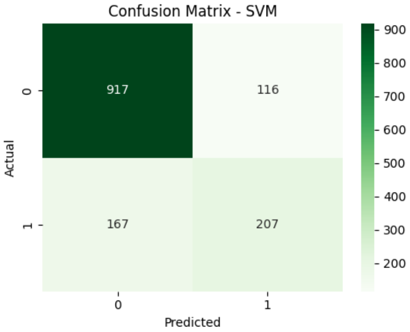

Fig. 11: Confusion Matrix: SVM (RBF) (After Tuning)

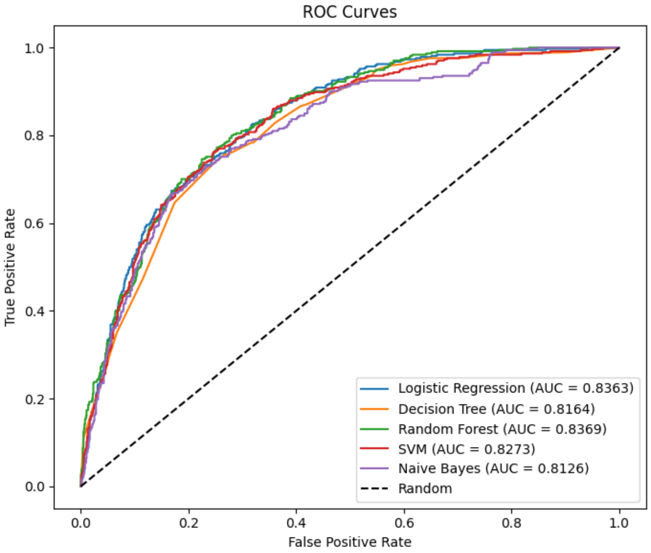

Fig. 12: ROC curves and AUC values for tuned models. Random Forest and Logistic Regression lead slightly with AUC 0.84.