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.

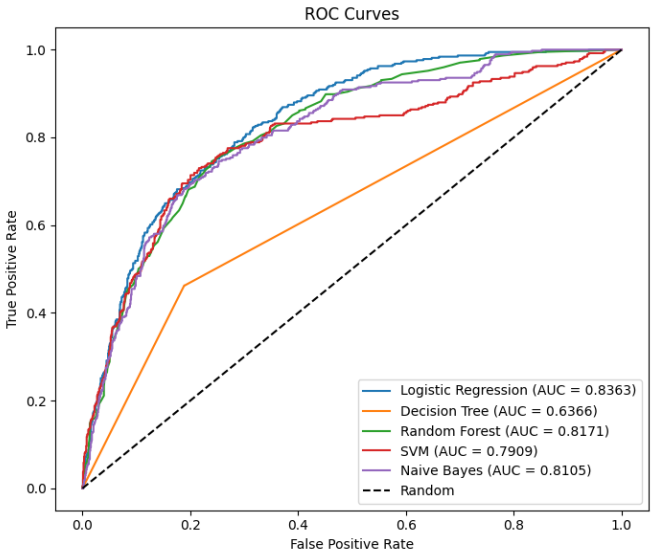

Fig. 1: ROC Curves for all models (before tuning).

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 post-tuning performance, Figures 812 display confusion matrices for all tuned classifiers.

Each matrix shows how true labels (rows) compare with predictions (columns):

- **Logistic Regression:** Improved detection of churners with balanced false negatives.
- **Decision Tree:** Cleaner class separation after pruning.
- **Random Forest:** Best accuracy and recall, minimal false positives.
- **Naive Bayes:** High recall but overpredicts churn (many FP), typical for independent-feature models.
- **SVM:** Strong precision with reduced false negatives due to tuned kernel parameters.

E. ROC Curves and AUC Analysis

Figure 13 shows ROC curves summarizing discrimination ability between churners and non-churners. Random Forest and Logistic Regression dominate with AUC around 0.84, followed by SVM (0.827), Decision Tree (0.816), and Naive Bayes (0.813). This demonstrates consistent performance gains after tuning.

F. Performance Summary

Random Forest yields the highest accuracy and recall, making it ideal for minimizing customer loss. Logistic Regression offers interpretability with nearly equal AUC. SVM provides strong margin separation, and Naive Bayes serves as a fast, explainable benchmark.

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

G. Business Interpretation

- Customers on month-to-month contracts are more likely to churn.
- High monthly charges correspond to elevated churn risk.
- Longer-tenure clients display strong retention probability.
- Models can automate early-warning churn alerts for targeted retention campaigns.

VI. Discussion

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

Future work includes exploring ensemble stacking, Gradient Boosting (XGBoost, LightGBM), model calibration, and explainable AI for business interpretability.

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 significantly enhanced 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 guidance and constructive 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.

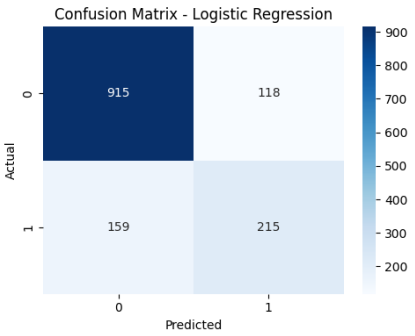

Fig. 2: Confusion Matrix: Logistic Regression

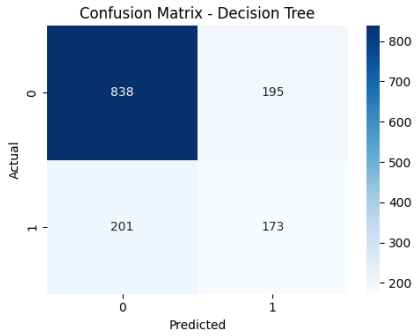

Fig. 3: Confusion Matrix: Decision Tree

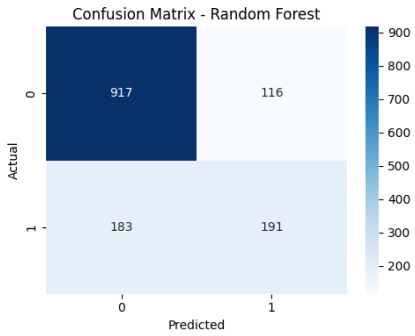

Fig. 4: Confusion Matrix: Random Forest

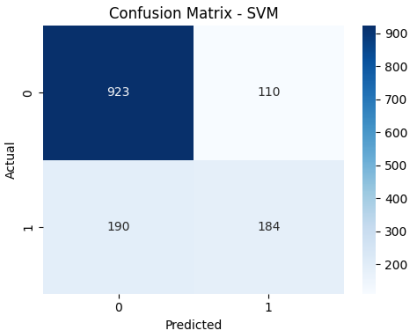

Fig. 5: Confusion Matrix: SVM (RBF)

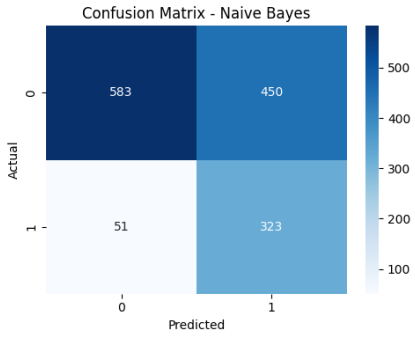

Fig. 6: Confusion Matrix: Naive Bayes (Before Tuning)

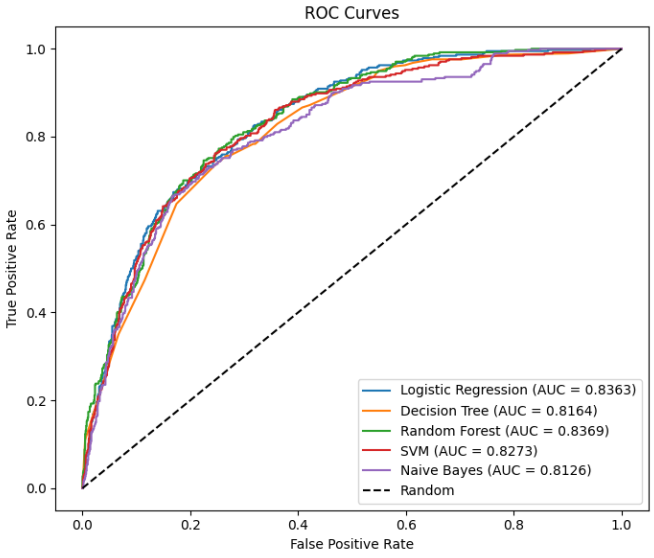

Fig. 7: ROC Curves for all models after hyperparameter tuning.

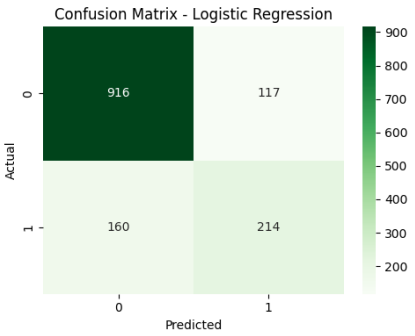

Fig. 8: Confusion Matrix: Logistic Regression (After Tuning)

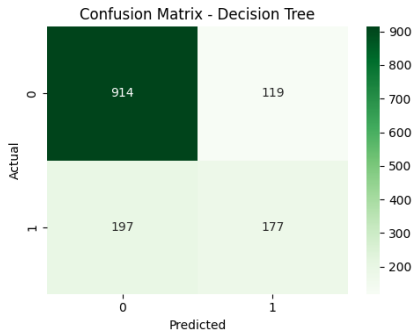

Fig. 9: Confusion Matrix: Decision Tree (After Tuning)

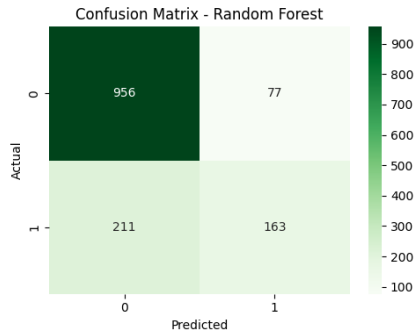

Fig. 10: Confusion Matrix: Random Forest (After Tuning)

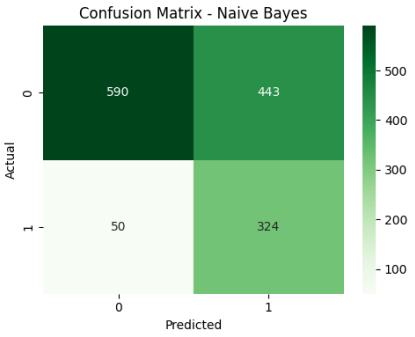

Fig. 11: Confusion Matrix: Naive Bayes (After Tuning)

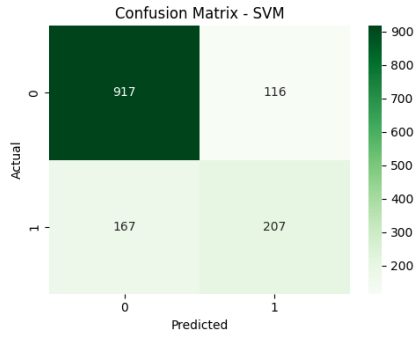

Fig. 12: Confusion Matrix: SVM (RBF) (After Tuning)

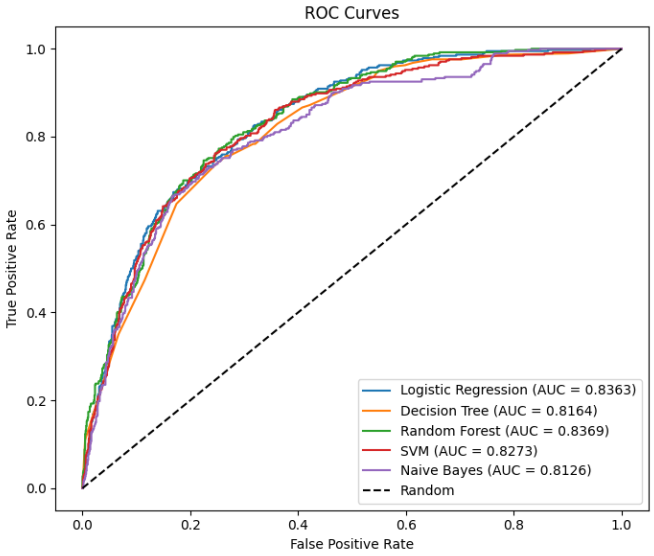

Fig. 13: ROC Curves (After Tuning): Random Forest and Logistic Regression achieve AUC 0.84.