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, 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.

Index Terms—Customer Churn, Classification, Data Mining, Machine Learning, Random Forest, 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; (iii) insights into the most influential predictors of churn.

II. Related Work

Churn prediction has been explored using classical statistical models and modern machine learning methods. Recent literature highlights the effectiveness of tree ensembles and cost-sensitive learning in handling class imbalance and heterogeneous predictor types. We focus on transparent baselines frequently reported in prior work and discuss extensions in Section VI.

III. Dataset

We use the Telco Customer Churn dataset (Kaggle), which contains 7,043 records with demographic attributes, service usage, contract type, payment method, and a binary churn label. The data includes a mix of numerical and categorical variables, motivating careful preprocessing and encoding. Missing values are handled using simple imputation strategies, and categorical features are one-hot encoded.

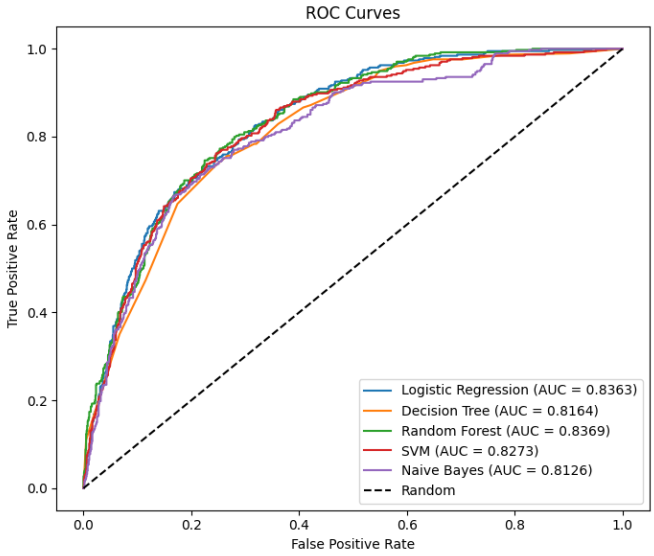

Fig. 1: ROC curves for all models.

IV. Methodology

Our pipeline follows a standard supervised learning workflow:

1) Data cleaning and preprocessing: type casting, missing value handling, outlier checks.
2) EDA: distributions, correlation heatmaps, churn rate by feature groups.
3) Feature engineering: one-hot encoding of categoricals; scaling of numeric features where required.
4) Modeling: Logistic Regression, Decision Tree, Random Forest, and SVM (RBF).
5) Evaluation: stratified train–test split; metrics include accuracy, precision, recall, F1-score; confusion matrices.
6) Hyperparameter tuning: grid search with cross-validation for tree depth, number of trees, and SVM parameters.

V. Experiments and Results

Feature importance: Contract type, tenure, monthly charges, internet service type, and payment method emerged as influential variables. These align with domain intuition that longer contracts and stable billing reduce churn propensity.

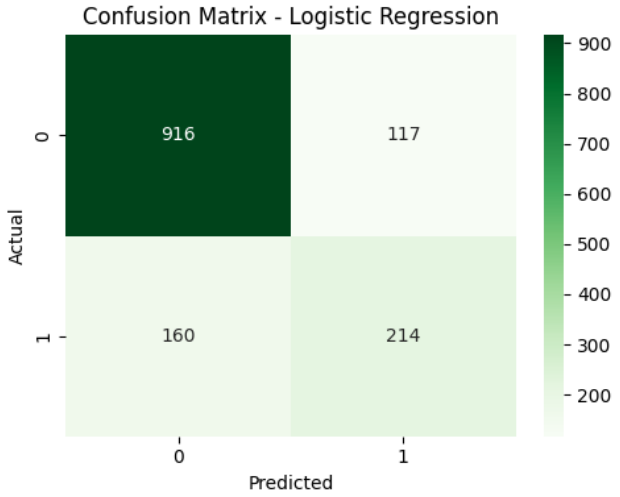

Fig. 2: Confusion Matrix: Logistic Regression

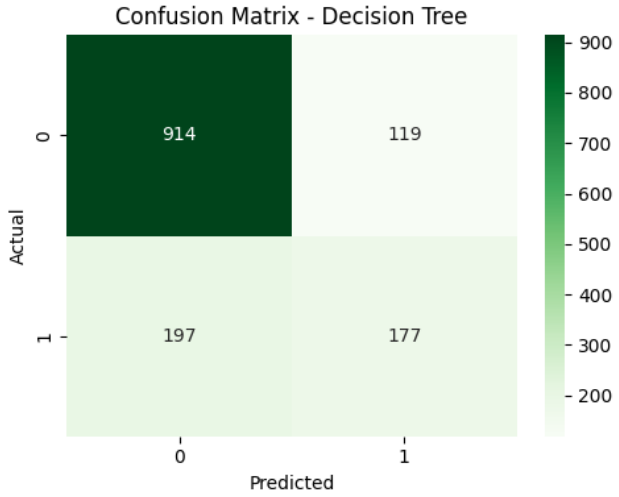

Fig. 3: Confusion Matrix: Decision Tree

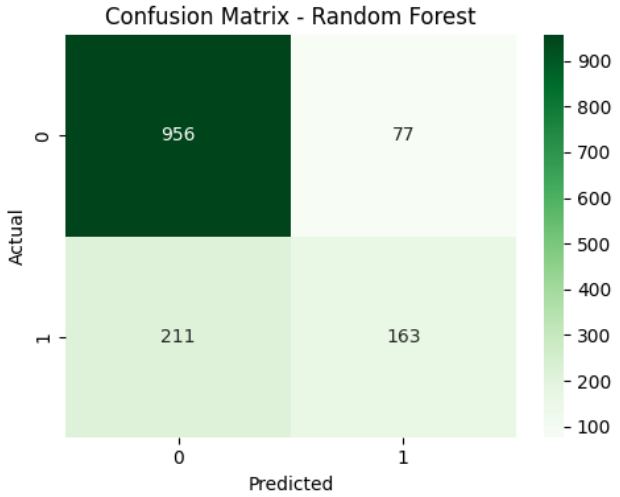

Fig. 4: Confusion Matrix: Random Forest

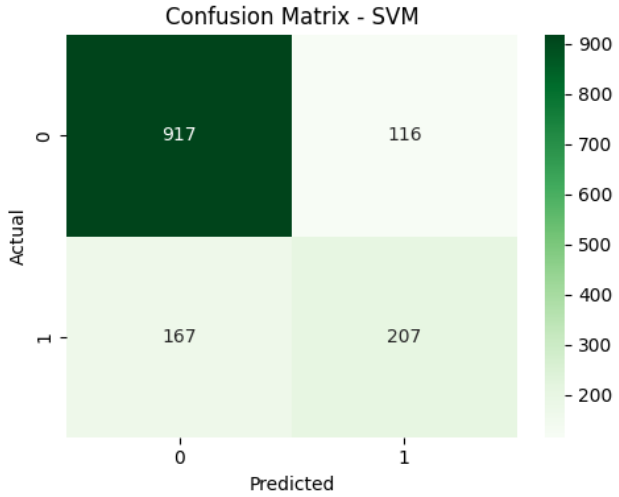

Fig. 5: Confusion Matrix: SVM (RBF)

TABLE I: Model performance summary.

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

VI. Discussion

Tree ensembles better capture non-linear interactions among heterogeneous features, which explains their superior empirical performance. However, high recall can come at the expense of precision, and model choices should be guided by the cost of false positives vs. false negatives in a given business context. Future improvements include gradient boosting (XGBoost/LightGBM), calibration, and cost-sensitive training.

VII. Conclusion

We presented a reproducible churn prediction pipeline on the Telco dataset and compared widely used classifiers. Random Forest offered the strongest balance across metrics for the churn class. The approach can be operationalized for early-warning retention, with care taken for model monitoring and drift.

Acknowledgments

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

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.