Telcom Recommendation System (TelRecom)

Veronica Obenewaa

90182026

Dr Isaac Nyantakyi

Machine Learning

24th December 2025

1.0 Project Overview

Customer churn remains a critical challenge in the telecommunications industry, where retaining existing customers is significantly more cost-effective than acquiring new ones. This project addresses the problem of predicting customer churn using machine learning techniques, with the objective of identifying customers who are likely to discontinue their services. The proposed solution integrates supervised learning models with systematic feature engineering and optimisation to build an accurate churn prediction system. Beyond prediction, the project also incorporates a recommendation pipeline that supports targeted retention strategies. The significance of this work lies in its ability to provide data-driven insights that enable telecom operators to proactively intervene, reduce revenue loss, and improve long-term customer lifetime value (CLTV).

2.0 Dataset Overview

The study utilises a telecommunications customer dataset comprising 7,043 observations with 33 columns, each representing an individual customer. The dataset includes demographic, geographic, service usage, and billing-related variables, along with churn indicators. Numeric features show that customers have an average tenure of 32.37 months, with values ranging from 0 to 72 months. Monthly charges vary widely, with a mean of 64.76 and a standard deviation of 30.09, indicating diverse pricing plans. The churn variable reveals that approximately 26.5% of customers have churned, highlighting a moderate class imbalance. Customer lifetime value (CLTV) has a mean of 4,400, underscoring the financial importance of retention. Geographic attributes such as latitude, longitude, and zip code provide spatial context but exhibit limited variability relative to behavioural features.

3.0 Exploratory data analysis

3.1 Churn Distribution

The distribution of the churn variable revealed a notable class imbalance, with approximately 73.5% of customers classified as non-churners and 26.5% identified as churners. This level of imbalance is typical in real-world churn prediction datasets, where the majority of customers tend to remain with the service provider.

3.2 Pairplot Analysis

The pair plot provides a comprehensive visualization of the relationships between customer tenure, monthly charges, total charges, and Customer Lifetime Value (CLTV), categorized by their churn status. A primary observation is the strong linear correlation between tenure months and total charges, where customers who do not churn tend to accumulate higher total charges over time, whereas churned customers are more densely clustered at the lower end of the tenure scale. The density plots on the diagonal reveal that churn is more prevalent among customers with high monthly charges, indicated by the blue peak near the 70 to 100 range, while non-churning customers show a more distributed billing profile. The visualization suggests that high monthly costs and short tenure are the most distinct indicators of customer churn.

3.3 Correlation analysis

Correlation analysis revealed moderate to strong relationships among several key features, indicating varying degrees of shared information. Tenure Months and Total Charges exhibited the strongest positive correlation ($r = 0.83$), suggesting substantial redundancy, as total charges naturally accumulate with longer customer tenure. A moderate positive correlation was observed between Churn Score and Churn Value ($r = 0.66$), indicating that higher predicted churn risk aligns well with actual churn outcomes. Similarly, Total Charges and Monthly Charges demonstrated a moderate correlation ($r = 0.65$), reflecting the influence of recurring billing on accumulated charges over time. In contrast, Monthly Charges and Customer Lifetime Value (CLTV) showed a weak correlation ($r = 0.10$), suggesting that higher monthly fees do not necessarily translate into greater long-term customer value. Notably, Churn Value and Tenure Months exhibited a moderate negative correlation ($r = -0.35$), indicating that customers with shorter tenure are more likely to churn

4.0 Data Preprocessing

4.1 Missing values and imputation

Missing values were identified in the Churn Reason and Total Charges variables. For the Churn Reason feature, missing entries were imputed with the category “Not Churned”, as the absence of a recorded churn reason logically corresponds to customers who did not churn. For the Total Charges variable, missing values were imputed using the median of the observed total charges.

Median imputation was chosen due to its robustness to outliers and skewed distributions, ensuring that extreme values did not disproportionately influence the imputed results.

4.2 Feature importance

Feature engineering was employed to enhance predictive performance, capture non-linear customer behaviours, and improve model interpretability. Tenure-based features, including tenure groups and early churn indicators, were introduced to reflect customer lifecycle effects, while spending-related features such as average monthly spend, price per service, and tenure-to-charge ratios captured perceived value and cost sensitivity. Service aggregation features, including total services, missing services count, and service engagement score, were designed to quantify customer engagement and switching costs. Contract and payment-based indicators, such as long-term contracts, high-charge month-to-month contracts, and risky payment methods, encoded behavioural risk patterns commonly associated with churn. Demographic interaction features, including senior-related indicators, improved sensitivity to vulnerable customer segments. Value-oriented features such as CLTV per month aligned churn prediction with business impact. These engineered features enabled ensemble models to learn meaningful interactions, and improve recall, stability, and overall model generalisation.

4.3 Encoding categorical variables

The dataset contained a total of 23 categorical features, which were transformed into numerical representations using a hybrid encoding strategy. Binary and low-cardinality categorical variables, such as gender, senior citizen status, and dependents, etc, were encoded using label encoding to produce concise 0/1 representations. In contrast, nominal categorical variables with multiple unordered classes, including contract type, payment method, and internet service, were encoded using one-hot encoding. All resulting dummy variables were explicitly converted to integer format to ensure numerical consistency and compatibility across machine learning models.

5.0 Model Training

5.1 Data split and scaling

The dataset was partitioned into training and testing subsets using an 80:20 split. Following the split, feature scaling was applied to the numerical variables to standardise their ranges and prevent features with larger magnitudes from disproportionately influencing the learning process

5.2 Models Trained

5.2.1 Logistic Regression

Logistic Regression was used as a baseline model to provide a transparent and interpretable reference for churn prediction. Exploratory data analysis revealed clear linear tendencies between key variables such as tenure and churn, particularly the inverse relationship indicating higher churn rates among low-tenure customers. Logistic Regression allows these linear effects to be directly quantified through model coefficients, enabling interpretability and serving as a benchmark against which more complex non-linear models can be evaluated.

5.2.2 Random Forest

Random Forest was selected based on EDA findings that indicated non-linear relationships and interaction effects among features, especially between tenure, monthly charges, and service usage. Pair plot visualisations showed that churned customers cluster at combinations of short tenure and high monthly charges, patterns that are difficult to capture with linear models. Random Forest effectively handles such non-linearities and correlated features, such as the strong correlation observed between tenure months and total charges, while reducing overfitting through ensemble averaging.

5.2.3XGBoost

XGBoost was employed due to its strong ability to model complex decision boundaries and handle structured tabular data with correlated predictors. EDA revealed moderate to strong feature correlations and class imbalance in the churn variable, conditions under which gradient boosting methods perform particularly well. XGBoost's regularization mechanisms and class weighting capabilities help mitigate overfitting and improve churn detection, especially for high-risk customer segments characterised by high charges and short tenure.

5.2.4 LightGBM

LightGBM was chosen as an efficient gradient boosting alternative capable of learning subtle non-linear patterns identified during EDA. The analysis showed overlapping distributions between churned and non-churned customers across several features, requiring models that can capture fine-grained splits and feature interactions. LightGBM's leaf-wise growth strategy enables faster convergence and improved scalability.

5.3 Handling Class Imbalance

SMOTE–Tomek was employed to address class imbalance in the churn prediction task, where the minority churn class was underrepresented in the training data. The Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic minority class samples to improve the model's ability to learn churn-related patterns, while Tomek links remove ambiguous and overlapping instances that lie near class boundaries. This combined approach enhances class separability, reduces noise, and mitigates model bias toward the majority class

5.4 Hyperparameter tuning

Hyperparameter tuning was applied because the initial models were not performing as expected, particularly in balancing churn detection and false positives. Using RandomizedSearchCV, the pipeline systematically explored combinations of hyperparameters for Logistic Regression, Random Forest, XGBoost, and LightGBM, optimizing for F1-score to achieve a balance between precision and recall. Parameters such as regularization strength, tree depth, number of estimators, and learning rate were tuned across cross-validation folds, and parallel processing (n_jobs=-1) was used for efficiency. The tuned models were then evaluated with threshold optimization, accuracy, AUC, classification report (precision, recall, F1 score), and confusion matrices to ensure improved predictive performance and more effective identification of at-risk customers.

6.0 Recommendation Pipeline

The customer churn prediction and plan recommendation pipeline integrates churn risk assessment, usage-based clustering, and personalized plan suggestions into a unified workflow. Initially, customers are segmented using K-Means clustering based on key usage metrics such as tenure, monthly charges, total services, and encoded categorical features like streaming, internet, and multi-line options. The optimal number of clusters is determined through inertia and silhouette analysis with silhouette score of 0.3735, and each cluster is characterized by size, average tenure,

service usage, and churn rate. Each cluster is then mapped to a tailored telecom plan, ranging from basic low-cost options for light or budget-conscious users to premium or family-oriented plans for high-usage or multi-line households. The pipeline further incorporates a pre-trained churn prediction model to classify customers into high- and low-risk categories. High-risk customers are clustered according to usage patterns and assigned plans aimed at churn prevention, while low-risk customers retain their current plans. The resulting output provides customer-level recommendations including churn probability, risk category, cluster assignment, recommended plan, and rationale. This approach enables data-driven retention strategies, targeted marketing, and scalable, automated plan optimization while ensuring reliability through clustering metrics and churn model performance evaluation.

7.0 Model Evaluation

7.1 Evaluation Before SMOTE

The evaluation results indicate that all models achieved comparable overall performance, with accuracy ranging between 75.5% and 79.1% and ROC–AUC values consistently above 0.84, demonstrating good discriminatory ability. Logistic Regression achieved the highest recall for churners (0.80), indicating strong sensitivity in identifying customers at risk of churn, but at the expense of lower precision, resulting in more false positives. Random Forest produced the highest overall accuracy (0.79) and the most balanced precision–recall trade-off for the churn class, reflecting stable and consistent classification performance. XGBoost also demonstrated strong recall (0.74) and competitive AUC, making it effective for identifying churners, though with moderate precision. LightGBM achieved performance comparable to Random Forest, with slightly lower recall but improved precision relative to Logistic Regression and XGBoost. Overall, ensemble models outperformed the baseline Logistic Regression in balancing churn detection and false alarms, while Logistic Regression remained valuable where maximising churn recall is prioritised.

7.2 Evaluation After SMOTE And Hyperparameter Tuning

The evaluation results after applying hyperparameter tuning and SMOTE and Tomek links show that all models improved their ability to detect churners, with accuracy ranging from 76.2% to 78.9% and ROC–AUC consistently above 0.84, indicating strong discriminatory power. Logistic

Regression achieved the highest precision (0.58) among all models while maintaining solid recall (0.71), reflecting fewer false positives compared to before oversampling. Random Forest showed the highest recall (0.80) after threshold adjustment, demonstrating strong sensitivity in identifying churners, although with moderate precision (0.53). XGBoost achieved the best overall F1-score (0.65) with balanced precision (0.57) and recall (0.76), making it the most effective model for capturing churn while controlling false alarms. LightGBM performed comparably to XGBoost, slightly lower in F1 (0.65) but maintaining high recall (0.77) and improved precision relative to baseline models. Overall, applying SMOTE with Tomek links enhanced recall across models, particularly benefiting ensemble methods, while XGBoost provided the best balance between churn detection and false positives, confirming its selection as the final model.

8.0 Conclusion

In conclusion, this project successfully developed a robust churn prediction and recommendation system for a telecommunications dataset, combining accurate predictive modeling with actionable retention strategies. Exploratory analysis identified key churn drivers, such as short tenure and high monthly charges, which guided feature engineering and model selection. Hyperparameter tuning along with SMOTE and Tomek links enhanced model sensitivity, with XGBoost achieving the best balance between detecting churners and minimizing false positives. The integrated recommendation pipeline further segments customers and provides personalized plan suggestions, enabling targeted interventions for high-risk customers. From a business perspective, the project demonstrates that data-driven insights can proactively identify customers at risk of churn and tailor retention offers, ultimately reducing revenue loss and improving long-term customer lifetime value.