# Predictive Analysis for Customer Churn

Aashish Singh[1], Prof. Soumya Jana[2]

***Abstract*— In the fiercely competitive telecommunications market, customer churn poses a significant challenge to companies' profitability and growth. Retaining existing customers is more cost-effective than acquiring new ones, making churn prediction a crucial aspect of customer relationship management (CRM). This project aims to develop a deep neural network model to predict customer churn in the telecom industry. The model utilizes a binary classification approach, leveraging customer data such as usage patterns, demographics, and service metrics to determine the likelihood of churn. The model incorporates multiple dense layers with leaky ReLU activations and employs the Adam optimizer and binary cross-entropy loss for enhanced performance. Early stopping is implemented to mitigate overfitting by monitoring validation loss and restoring the best model weights. The project's intermediate results demonstrate the model's learning capability, with training loss decreasing and accuracy steadily increasing over epochs. The model has achieved promising performance metrics, including an accuracy of 93%, a precision of 77%, a recall of 85%, and an F1 score of 0.81. Future steps involve analyzing feature impact, exploring dimensionality reduction, experimenting with different models, and further enhancing the model's accuracy. By accurately predicting churn, this project will empower businesses to implement proactive retention strategies, minimize customer loss, and foster long-term customer loyalty.**

## I. Introduction

Customer churn, the act of customers discontinuing their service with a company, poses a significant challenge across industries, particularly in the highly competitive telecommunications market. Acquiring new customers is more expensive than retaining existing ones, so churn prediction is crucial for telecom companies' profitability. This focus on customer retention aligns with strategies for maximizing profits and remaining competitive.

The telecommunications industry has experienced substantial transformations due to market liberalization, new services, and technological advancements. This intensified competition has amplified the importance of customer churn management. Retaining existing customers is more profitable than acquiring new ones, leading businesses to prioritize customer-centric targeted marketing strategies over mass marketing. Accurately predicting customer churn enables companies to implement proactive retention strategies, enhance customer satisfaction, and minimize revenue loss. Fortunately, telecom companies have access to vast amounts of customer data, including usage patterns, demographics, and service interactions. This data can be leveraged to identify patterns and build predictive models using machine learning (ML) and data mining techniques.

Various ML algorithms have been applied to customer churn prediction, including:
Decision trees, Support Vector Machines (SVMs), Neural networks [1], Ensemble methods like Random Forest [2] and Adaboost [3].

Regression models, particularly logistic regression, are popular due to their high reported accuracy and interpretability in identifying key churn drivers. However, the class imbalance problem, with a significantly smaller proportion of churners compared to non-churners, poses a challenge in churn prediction. Sampling techniques are often used to address this imbalance.

This project investigates the application of diverse ML algorithms to predict customer churn in the telecom industry, aiming to determine the most accurate and insightful model. It builds upon previous work showcasing the potential of a deep neural network and extends the analysis to include a wider range of models:
Neural Network, Logistic Regression, Random Forest,SVM, KNN, Naive Bayes, Linear Regression, LDA, Decision Tree

The project evaluates each model using accuracy, precision, recall, F1-score, and ROC-AUC to assess its predictive ability. The analysis compares the performance of these models, drawing insights into their strengths and limitations for churn prediction.

By systematically evaluating these models, the project aims to identify the most effective algorithm for churn prediction, enhance understanding of churn drivers, and contribute valuable insights to developing robust churn management strategies in the telecom industry.

## II. Literature Review

Customer churn is a significant issue for companies in the telecommunications industry, impacting revenue and highlighting the need for predictive models to enable proactive retention efforts. Researchers have investigated various approaches to churn prediction, including traditional statistical methods, machine learning techniques, and the use of big data platforms. While early studies explored statistical techniques, the focus has shifted toward more advanced methods like machine learning and data mining for improved accuracy and the ability to handle large datasets. Machine learning algorithms, including Decision Trees, Support Vector Machines, Neural Networks, Naive Bayes, Logistic Regression, ensemble methods (e.g., AdaBoost, Bagging), and evolutionary algorithms (e.g., Genetic Programming), have been widely investigated for their effectiveness in churn

[1]Department of Artificial Intelligence, Indian Institute of Technology Hyderabad, India.
[2]Department of Electrical Engineering, Indian Institute of Technology Hyderabad, India.

prediction. The choice of the most appropriate algorithm depends on factors such as data characteristics, desired interpretability, and performance metrics.

The sources emphasize the importance of feature engineering, which involves creating relevant features from raw data to enhance model performance. These features can be derived from various sources, such as call detail records (CDRs), customer demographics, billing information, service usage patterns, and social network interactions [4]. Additionally, feature selection techniques are employed to identify the most informative features, improving model accuracy and interpretability. Researchers commonly use techniques like Information Gain, Correlation Attribute Ranking Filter, and Principal Component Analysis (PCA) for feature selection.

A common challenge in churn prediction is the presence of imbalanced datasets, where non-churners significantly outnumber churners, potentially biasing model training. Techniques like undersampling, oversampling, or the use of algorithms less sensitive to imbalance are often employed to mitigate this issue. The selection of appropriate evaluation metrics is also crucial, as different metrics provide different insights into model performance. Commonly used metrics include accuracy, precision, recall, F1-score, and ROC AUC, with the choice depending on the specific goals of the prediction task. For instance, if minimizing false positives is crucial, precision becomes a key metric, while recall is prioritized when maximizing the identification of true churners.

Recent advancements in churn prediction include the integration of Social Network Analysis (SNA) [4] features derived from customer social networks, showing promise in improving accuracy by considering social influence on churn behavior. The exploration of real-time prediction using data streams and online learning algorithms is another emerging trend, enabling immediate intervention for customer retention. Researchers are also investigating hybrid approaches that combine different machine learning techniques or integrate statistical methods with machine learning to further enhance prediction accuracy.

## III. Materials and Methods

This section outlines the materials and methods employed to predict customer churn in the telecommunications industry. The study uses a dataset comprising various customer attributes to develop and evaluate different machine learning models for churn prediction.

### A. Data Description

The dataset under investigation is designed to predict customer churn using a binary classification approach. It encompasses a range of customer-related features categorized into behavioral, demographic, and service-related attributes.

A detailed description of these features and their corresponding values, as presented in the data sample, is provided below:

- **Call Failure** - This feature represents the number of times a customer experienced call failures.

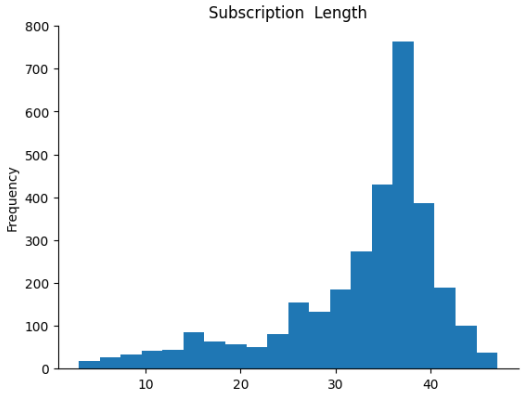


Fig. 1: Histogram showing relationship between Subscription Length and frequency.

- **Complains** - This feature indicates whether a customer has filed complaints. Values in the sample are binary, with 0 representing no complaints and 1 indicating a complaint.
- **Subscription Length** - This numerical feature represents the duration of a customer's subscription.
- **Charge Amount** - This feature indicates the amount charged to the customer for the service.
- **Seconds of Use** - This feature represents the total duration of service usage by the customer, measured in seconds.
- **Frequency of use** - This numerical feature indicates the frequency of service usage by a customer.
- **Frequency of SMS** - This feature reflects how often a customer uses SMS services.
- **Distinct Called Numbers** - This feature represents the number of unique phone numbers a customer contacted.
- **Age Group** - This categorical feature classifies customers into different age groups.
- **Tariff Plan** - This feature represents the specific tariff plan subscribed to by the customer.
- **Status** - This feature, with binary values, indicates the customer's current status.
- **Age** - This feature represents the customer's age in years.
- **Customer Value** - This feature quantifies the value of a customer, likely based on their revenue generation or profitability.
- **Churn** - The target variable, represented as a binary value (0 or 1), indicates whether a customer has churned. In the sample, 0 represents a non-churned customer, while 1 signifies a churned customer.

### B. Methodology

*1) Data Preprocessing:* The dataset was first imported from a CSV file containing features relevant to customer churn analysis. Key preprocessing steps included:

1) **Feature Selection:** Redundant features were removed to ensure the model focuses on relevant predictors for churn. For instance, features such as *Subscription*


Fig. 2: Flowchart that visualizes the process of applying the neural network for churn prediction.

*Length*, *Seconds of Use*, and *Frequency of Use* were retained while excluding others deemed irrelevant.

2) **Splitting Data:** The data was divided into features ($X$) and the target variable ($y$), where $y$ indicates whether a customer churned (1) or not (0).
3) **Train-Test Split:** Using an 80-20 split, the dataset was divided into training and testing sets to ensure proper validation. The split ensures unbiased performance evaluation on unseen data.
4) **Normalization:** To standardize the feature values, the `StandardScaler` was employed:

$$X' = \frac{X - \mu}{\sigma}$$

where $X$ is the feature value, $\mu$ is the mean, and $\sigma$ is the standard deviation.

*2) Model Implementation:* The project employed various machine learning models for churn prediction. Each model was tuned and evaluated using common performance metrics.

1) **Neural Network:** A feedforward neural network was implemented with the following architecture:
   - Input Layer: Number of features.
   - Hidden Layers: 4 layers with 8 neurons each and Leaky ReLU activation.
   - Output Layer: 1 neuron with sigmoid activation to predict the probability of churn.

   The network was optimized using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy as the loss function:

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i))$$

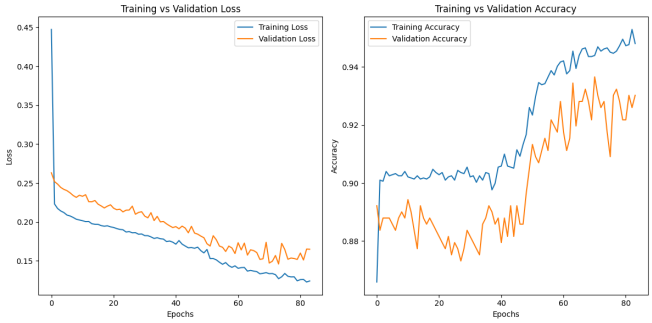


Fig. 3: (a) Model's training loss decreases over epochs, (b) The change in training and validation accuracy over time

2) **Logistic Regression:** A baseline model using logistic regression was implemented to establish a benchmark:

$$P(y=1|X) = \frac{1}{1+e^{-(wX+b)}}$$

3) **Random Forest:** Utilized 100 decision trees, employing the Gini impurity criterion:

$$G = 1 - \sum_{i=1}^{k} p_i^2$$

where $p_i$ is the proportion of samples belonging to class $i$.
4) **Other Models:** Support Vector Machines (SVM) with Radial Basis Function kernel, K-Nearest Neighbors (KNN), Naive Bayes, Linear Discriminant Analysis (LDA), and Decision Tree classifiers were also implemented for comparison.

*3) Model Evaluation:* Performance metrics used to evaluate the models include:

- **Accuracy:** Fraction of correctly classified instances:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

- **Precision:** Correctly predicted positive cases over total predicted positives:

$$\text{Precision} = \frac{TP}{TP+FP}$$

- **Recall:** Correctly predicted positive cases over total actual positives:

$$\text{Recall} = \frac{TP}{TP+FN}$$

- **F1-Score:** Harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **ROC-AUC:** Area under the ROC curve to measure separability between classes.

*4) Error Analysis:* The error rate for each model was calculated as:

$$\text{Error Rate} = \frac{FP+FN}{TP+TN+FP+FN}$$

Confusion matrices were used to visualize classification performance, helping identify common misclassifications.

*5) Comparison:* The models' performances were compared using visualizations. Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were plotted to identify the best-performing algorithm for churn prediction. Random Forest achieved the highest accuracy (95%) and ROC-AUC (98%), indicating its superior predictive capability for this dataset.

## IV. Experimental Results

The experimental results provide a comprehensive comparison of the performance of various machine learning models on the customer churn dataset. Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were used to evaluate the models, highlighting their strengths and limitations in addressing the churn prediction problem.

*1) Performance Evaluation:* Among the tested models, the Random Forest classifier emerged as the best-performing model, achieving an accuracy of 95% and an ROC-AUC score of 98%. These results highlight its ability to handle the dataset's features effectively and differentiate between churned and non-churned customers with high precision. The F1-score of 0.84 indicates a balance between precision and recall, making it well-suited for this binary classification task.

On the other hand, the Naive Bayes model exhibited the lowest performance, with an accuracy of 73%, a precision of 39%, and an F1-score of 0.54. While its recall was relatively high (93%), this came at the cost of a significant number of false positives, making it less reliable for this specific application. This result underscores the model's assumption of feature independence, which may not hold true in the context of customer churn data.

*2) Model Selection Considerations:* The choice of the Random Forest model is justified by its superior performance metrics and inherent advantages, such as its ability to handle feature interactions and avoid overfitting through ensemble learning. However, the complexity of the model comes at the cost of longer training times compared to simpler models like Logistic Regression, which achieved an accuracy of 88% and an ROC-AUC of 93%. Despite its lower overall performance, Logistic Regression remains a viable choice for scenarios where interpretability and computational efficiency are prioritized.

Neural Networks also performed well, achieving an accuracy of 93% and an ROC-AUC of 93%. While their flexibility allows for capturing complex relationships in data, they require careful tuning of hyperparameters and are computationally intensive. The Support Vector Machine (SVM) demonstrated comparable performance, with an accuracy of 91% and an ROC-AUC of 96%. Its ability to create non-linear decision boundaries makes it suitable for datasets with complex feature interactions, though it may struggle with scalability for larger datasets.

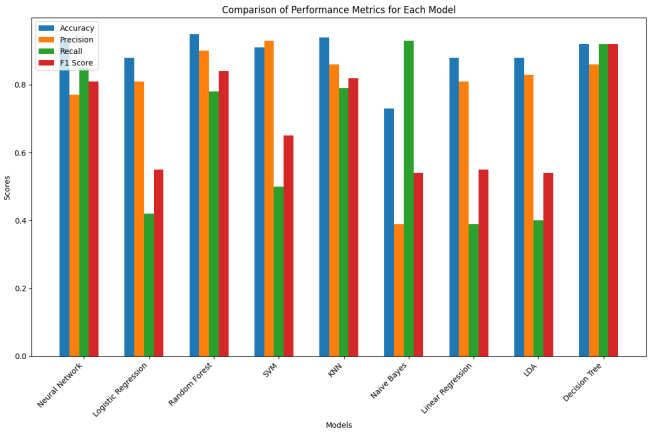


Fig. 4: Comparison between performance of different models

*3) Understanding Through Simplicity:* The problem chosen for this study is relatively less complex, as it allows for a clear and direct comparison of different models without the confounding effects of high-dimensional or highly imbalanced data. This simplicity facilitates understanding the strengths and limitations of each model, making it an ideal starting point for exploring machine learning techniques. It is noteworthy that for more complex real-world problems, model selection would require deeper consideration of factors such as feature engineering, domain-specific constraints, and computational resources.

*4) Conclusion on Model Usage:* In summary, the Random Forest model is recommended for customer churn prediction due to its robustness and high predictive power. Simpler models like Logistic Regression and Decision Trees may be considered for cases where interpretability or speed is critical, while more advanced models like Neural Networks or SVMs can be deployed when the dataset or problem complexity demands their capabilities. The results of this study emphasize the importance of tailoring model selection to the specific requirements and constraints of the problem at hand.

| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
|---|---|---|---|---|---|
| Neural Network | 0.93 | 0.77 | 0.85 | 0.81 | 0.93 |
| Logistic Regression | 0.88 | 0.81 | 0.42 | 0.55 | 0.93 |
| Random Forest | 0.95 | 0.90 | 0.78 | 0.84 | 0.98 |
| SVM | 0.91 | 0.93 | 0.50 | 0.65 | 0.96 |
| KNN | 0.94 | 0.86 | 0.79 | 0.82 | 0.97 |
| Naive Bayes | 0.73 | 0.39 | 0.93 | 0.54 | 0.91 |
| Linear Regression | 0.88 | 0.81 | 0.39 | 0.55 | 0.93 |
| LDA | 0.88 | 0.83 | 0.40 | 0.54 | 0.93 |
| Decision Tree | 0.92 | 0.86 | 0.92 | 0.92 | 0.86 |

TABLE I: Performance Metrics for Different Models

## V. Conclusion

This study aimed to explore the application of various machine learning models for predicting customer churn in the telecom sector. By experimenting with a range of models, including Neural Networks, Logistic Regression, Random

Forest, Support Vector Machines, and others, this project provided valuable insights into their performance. The results demonstrated that the Random Forest classifier outperformed other models in terms of accuracy, ROC-AUC, and overall robustness, making it a strong candidate for churn prediction tasks.

The models' performance highlighted the trade-offs between model complexity, computational cost, and accuracy. Simpler models like Logistic Regression offered interpretability and efficiency, while more complex models, such as Random Forest and Neural Networks, demonstrated their ability to capture more intricate patterns in data. This study emphasized the importance of aligning model selection with the specific requirements and constraints of the problem at hand.

## VI. Discussion

*1) Key Learnings:* The project provided several key takeaways:

- **Data Preprocessing:** Effective data preprocessing, including feature selection and normalization, played a critical role in model performance. Proper handling of features ensured that the models could learn relevant patterns in the data.
- **Model Trade-offs:** Each model demonstrated distinct strengths and limitations. While Random Forest excelled in accuracy and robustness, simpler models like Logistic Regression provided an advantage in terms of computational efficiency and interpretability.
- **Evaluation Metrics:** Employing multiple evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC allowed for a comprehensive assessment of each model's performance.
- **Hyperparameter Tuning:** The performance of more complex models, such as Neural Networks and Random Forest, heavily depended on hyperparameter tuning, underscoring the importance of optimization for achieving the best results.

*2) Practical Implications:* The findings of this study can be directly applied to real-world churn prediction scenarios in the telecom industry. The Random Forest model, with its superior ability to capture feature interactions and deliver high accuracy, is well-suited for predicting customer churn. However, for organizations with limited computational resources, simpler models such as Logistic Regression may provide a balance between performance and efficiency. These models can be particularly useful for organizations that require faster execution times and clearer interpretability.

*3) Limitations and Future Work:* The dataset used in this study, though detailed, is complex due to the wide range of features involved in predicting customer churn. Real-world datasets often present challenges such as high dimensionality, missing values, and class imbalance, all of which need to be addressed for more accurate modeling. Future work could involve:

- Applying these models to larger and more diverse datasets to test their scalability and robustness in different contexts.
- Investigating advanced techniques such as ensemble methods, deep learning architectures, or automated machine learning (AutoML) to further enhance predictive performance.
- Exploring explainability and interpretability techniques for complex models to ensure that the results are actionable and aligned with business needs.

In conclusion, this project provided a thorough comparative analysis of several machine learning models and highlighted their effectiveness in the task of customer churn prediction. The insights gained offer a solid foundation for further research and the practical application of these models in real-world customer retention strategies.

## References


[1] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Systems with Applications*, 2012.

[2] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE Access*, 2019.

[3] N. Lu, H. Lin, J. Lu, and G. Zhang, "A customer churn prediction model in telecom industry using boosting," *IEEE Transactions on Industrial Informatics*, 2014.

[4] A. Ahmad, K. J. Ahmad, and K. Jabbar, "Customer churn prediction model based on mobile social network analysis using big data platform," *Journal of Big Data*, 2019.