# Customer Churn Prediction Using Random Forest, Decision Tree

Sravani, Divya, Lahari, Sruthi

*Abstract*—***Customer churn poses a significant challenge for many industries, particularly those with subscription-based models such as telecommunications, banking, and streaming services. With increasing competition, the ability to retain customers becomes critical for sustained growth. Machine learning techniques have emerged as powerful tools in predicting customer churn by leveraging vast amounts of data, from demographic details to service usage patterns. This project focuses on the development and evaluation of several machine learning models aimed at predicting customer churn using a telecommunications dataset. The dataset was preprocessed through feature encoding, scaling, and addressing data imbalance using the SMOTEENN technique to ensure an even distribution of churned and non-churned customers in the training set.***

***Several models were trained and evaluated, including Logistic Regression, Support Vector Machines, K-Nearest Neighbors, and Random Forest Classifiers. Each model's performance was assessed based on accuracy, precision, recall, and F1 score, with Random Forest emerging as the best-performing model, achieving an accuracy of 94.27%. To explore the impact of feature reduction, Principal Component Analysis (PCA) was applied, but it did not enhance the model's performance. The final model was pickled for future deployment and integration into an API, making it accessible for real-time churn predictions. By providing actionable insights into which customers are most likely to churn, this model can be integrated into retention strategies, enabling targeted interventions that can reduce churn and increase customer loyalty.***

## I. Introduction

Customer churn is the term used to describe the process of customers discontinuing their business relationship with a company, whether by canceling a subscription, switching service providers, or opting out of a contract. The ability to predict which customers are likely to churn before they actually leave is crucial in highly competitive markets, where acquiring new customers is often more expensive than retaining existing ones. Industries such as telecommunications, financial services, retail, and subscription-based businesses are particularly susceptible to the impacts of churn[1]. As businesses grow, the complexity of analyzing customer behavior and identifying churn patterns increases exponentially. This is where machine learning (ML) models offer a competitive advantage.

The motivation behind predicting churn is twofold: it not only allows businesses to take preemptive actions to retain customers but also provides insights into customer behavior and preferences. With a well-trained churn prediction model, businesses can reduce their customer acquisition costs by improving customer retention. The primary objective of this project is to build a predictive model using historical data to identify customers at risk of churning. By analyzing customer data related to demographic factors, service usage, and billing information, we aim to construct a model that can provide accurate and timely predictions of churn.

The project follows a structured approach, starting with data preprocessing, exploratory data analysis (EDA), and feature engineering. Various machine learning models are trained, and their performance is evaluated using metrics such as accuracy, precision, recall, and F1 score. The project also addresses class imbalance, which is a common challenge in churn datasets, where the number of customers who churn is typically much smaller than those who do not. By applying techniques such as SMOTEENN, we can balance the dataset to improve the model's ability to detect churn. The ultimate goal is to create a robust model that can be deployed in real-time environments, providing businesses with a tool to reduce churn and improve customer satisfaction.

## II. LITERATURE REVIEW

The prediction of customer churn has been an area of active research across multiple domains. The importance of churn prediction lies in its ability to help businesses anticipate and reduce customer attrition, which can significantly impact revenue. The literature on churn prediction highlights several machine learning techniques and strategies that have been successfully applied to this problem.

Early work in churn prediction primarily focused on statistical methods such as Logistic Regression (LR) and Decision Trees (DT). These methods were simple to implement and offered interpretable results. However, as the volume of data increased and the complexity of customer behavior grew, more sophisticated methods were required. Ensemble methods such as Random Forest (RF) and Gradient Boosting Machines (GBM) became popular due to their ability to handle large datasets with high-dimensional features and their robustness to overfitting[1]. In particular, Random Forest, an ensemble learning method that operates by constructing multiple decision trees, has been shown to perform well in classification tasks, including churn prediction.

One of the major challenges in churn prediction is the imbalance in the dataset, where the proportion of churned customers is often much smaller than non-churned customers. To address this, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) have been developed.

https://doi.org/10.17562/PB-62-2

SMOTE works by generating synthetic samples for the minority class, thereby balancing the class distribution. Other variations, such as SMOTEENN, combine over-sampling with under-sampling of the majority class to create a more balanced and informative dataset. Studies have shown that using these techniques can significantly improve the performance of machine learning models in churn prediction.

Dimensionality reduction is another important aspect in churn prediction. High-dimensional data can lead to overfitting, where the model learns patterns that are specific to the training data but do not generalize to new data [1]. Principal Component Analysis (PCA) is a widely used technique for reducing dimensionality by transforming the features into a lower-dimensional space while retaining most of the variance in the data. However, the effectiveness of PCA varies depending on the dataset and model being used. In some cases, it can improve model performance, while in others, it may not offer significant benefits.

In this project, we explore several of these techniques, including Random Forest, SMOTEENN, and PCA, to develop an optimal model for churn prediction. The insights gained from this study contribute to the growing body of research on churn prediction, particularly in the context of handling imbalanced datasets and leveraging ensemble learning methods.

## III. METHODOLOGY

### A.Data Collection and Preprocessing

The dataset used for this project comes from a telecommunications company and contains a variety of features related to customer demographics, service usage, and billing information. Some of the key features include `gender`, `tenure`, `monthly charges`, `internet service`, and `contract type`. The target variable is `churn`, which indicates whether or not a customer has discontinued their service.

Before applying machine learning models, the dataset must be preprocessed to ensure that it is clean and suitable for analysis. Preprocessing steps include handling missing data, encoding categorical variables, and scaling numerical features. For example, categorical variables such as `gender` and `internet service` are encoded using one-hot encoding, which transforms them into binary vectors. Numerical features such as `tenure` and `monthly charges` are standardized to ensure that they are on a similar scale [2].

### B.Exploratory Data Analysis (EDA)

Exploratory data analysis is performed to gain insights into the dataset and identify any patterns or trends that may be useful for predicting churn. During EDA, we examine the distribution of key features and their relationships with the target variable, `churn`. For example, customers with shorter tenure and those with month-to-month contracts tend to have higher churn rates, while customers with longer contracts and higher service usage are less likely to churn.

Correlation analysis is also conducted to identify relationships between features. Features that are highly correlated with each other may be redundant and can be removed to simplify the model. Visualization techniques such as histograms, bar charts, and box plots are used to present the findings from EDA.

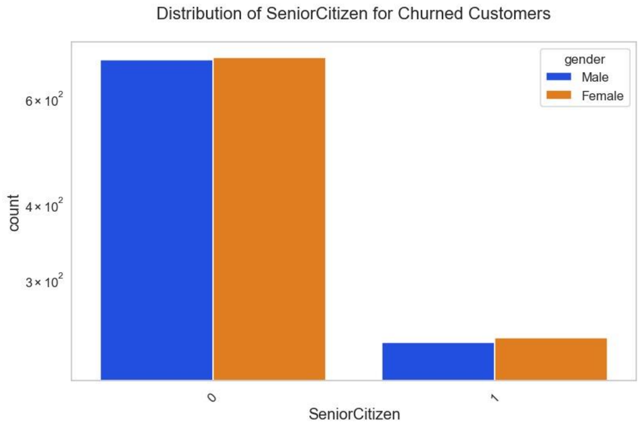


Fig.1.EDA Picture

### C. Model Selection and Training

Several machine learning models are tested in this project, including Logistic Regression, K-Nearest Neighbors, Support Vector Machines, and Random Forest Classifiers. Each model has its strengths and weaknesses, and the goal is to select the model that performs best on the test data. Cross-validation is used to evaluate the models' performance and avoid overfitting.

The Random Forest classifier emerged as the best-performing model in this project. It is an ensemble method that creates multiple decision trees during training and averages their predictions to make a final decision. Random Forests are known for their robustness and ability to handle both categorical and continuous features.

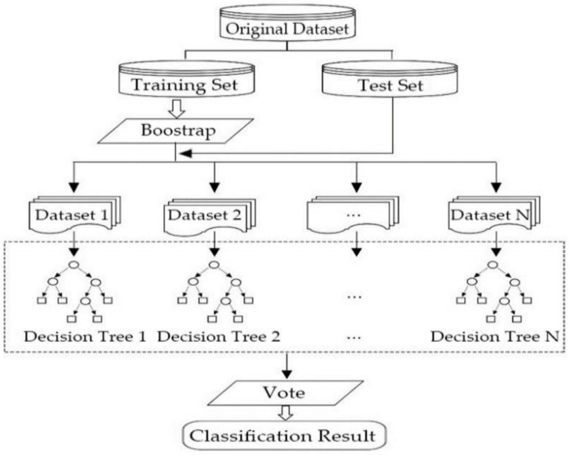


Fig.2.Model preparation

D. Addressing Data Imbalance

One of the main challenges in this project is the class imbalance in the dataset. The number of customers who churn is much smaller than those who do not, which can lead to biased models that perform poorly in identifying churned customers. To address this, we use the SMOTEENN technique, which combines SMOTE (Synthetic Minority Over-sampling Technique) with ENN (Edited Nearest Neighbors). SMOTE generates synthetic samples for the minority class (churned customers), while ENN removes noisy samples from the majority class. This results in a more balanced and representative dataset for training the model.

SMOTE formula:

The formula for generating a synthetic point Xsynthetic is:

$$Xsynthetic=Xi+\lambda\times(Xnn-Xi)$$

- Xi be a data point from the minority class.
- Xnn be one of the k-nearest neighbors of Xi
- λ be a random number in the range [0, 1].

ENN (Edited Nearest Neighbors) Mathematical Formula:

The removal condition in ENN is as follows:

If C^j!=Cj, then Xj is removed.

- Xj be a data point from the dataset.
- Knn(xj) be the k-nearest neighbors of Xj.
- Cj be the class label of Xj.
- C^j be the predicted class of Xj, which is determined by majority voting among Knn(Xj).

E. Dimensionality Reduction

Dimensionality reduction techniques such as PCA (Principal Component Analysis) are applied to reduce the number of features in the dataset. By transforming the features into a lower-dimensional space, we can reduce the risk of overfitting and improve model interpretability. However, in this project, PCA did not lead to significant improvements in model performance, so it was not included in the final model

## IV. RESULTS AND DISCUSSION

The results of the machine learning models developed in this project were evaluated using a variety of metrics, including accuracy, precision, recall, and F1-score. Each metric provides a different perspective on the model's performance, allowing for a more comprehensive assessment [3]. The Random Forest model, trained with the SMOTEENN technique to handle class imbalance, achieved the highest accuracy of 94.27%. This indicates that the model was able to correctly predict whether a customer would churn or not in 94.27% of cases on the test set.

A.Model Evaluation Metrics

- Accuracy: The overall accuracy of the Random Forest model was 94.27%, indicating that a large majority of the predictions were correct. However, accuracy alone can be misleading in imbalanced datasets like churn prediction, where the number of non-churned customers far outweighs the churned ones.
- Precision: The precision of the model was 0.72, which indicates that when the model predicted a customer would churn, it was correct 72% of the time. Precision is an important metric when the cost of a false positive is high (e.g., unnecessarily targeting customers who were not likely to churn).
- Recall: The recall of the model was 0.81, meaning that it successfully identified 81% of the actual churned customers. Recall is particularly important when the goal is to capture as many of the true positive cases (churned customers) as possible, even if some false positives are allowed.
- F1-Score: The F1-score, which balances precision and recall, was 0.77. This is a good indicator of the model's overall effectiveness, as it takes both false positives and false negatives into account.

B. Confusion Matrix

The confusion matrix for the Random Forest model provided additional insights into its performance. It showed that the model was able to correctly classify most of the non-churned customers, as expected in an imbalanced dataset. However, it also successfully identified a large proportion of the churned customers, which is crucial for the purpose of the project. The confusion matrix revealed that there were still some false negatives (i.e., customers who churned but were predicted not to), which could be further minimized in future iterations of the model by improving recall.

### C. Effect of SMOTEENN

The SMOTEENN technique had a significant impact on the model's ability to correctly identify churned customers. Without addressing the class imbalance, the model would likely have been biased towards predicting non-churned customers, as they represent the majority class. By using SMOTEENN, the minority class (churned customers) was better represented in the training data, allowing the model to learn patterns associated with churn more effectively. The improvement in recall and F1-score highlights the importance of using appropriate techniques to handle imbalanced datasets.

### D. Impact of PCA on Model Performance

Dimensionality reduction using PCA was tested to explore whether reducing the number of features would improve the model's performance or reduce overfitting. However, in this case, PCA did not lead to a significant improvement in model accuracy or other evaluation metrics [4]. This suggests that the original set of features, after preprocessing and feature engineering, was already well-optimized for the model. Additionally, the Random Forest model is inherently capable of handling high-dimensional data, which may explain why PCA did not provide any added benefit. While PCA can be useful in some scenarios, it is not always necessary, especially when using models like Random Forest that are robust to large feature sets.

### E. Limitations

Despite the strong performance of the Random Forest model, there are some limitations to the approach taken in this project. One limitation is that the dataset used for training may not capture all the factors that influence customer churn. For example, external factors such as market competition, economic conditions, or customer satisfaction with competitors' services were not included in the dataset. Incorporating such variables in future models could provide a more comprehensive view of customer behavior. Additionally, while SMOTEENN helped balance the dataset, it is still possible that the model may underperform in real-world settings where class distributions may differ.

## V. CONCLUSION AND FUTURE WORKS

This project successfully developed a machine learning model capable of predicting customer churn with high accuracy. The use of Random Forest, in combination with SMOTEENN to address class imbalance, resulted in a model that achieved strong performance across multiple evaluation metrics, including accuracy, precision, recall, and F1-score. The project demonstrates the importance of addressing imbalanced datasets and selecting appropriate machine learning techniques for classification tasks. By identifying customers at risk of churning, businesses can take targeted actions to retain them, thereby improving customer satisfaction and reducing customer acquisition costs.

### A.Conclusion

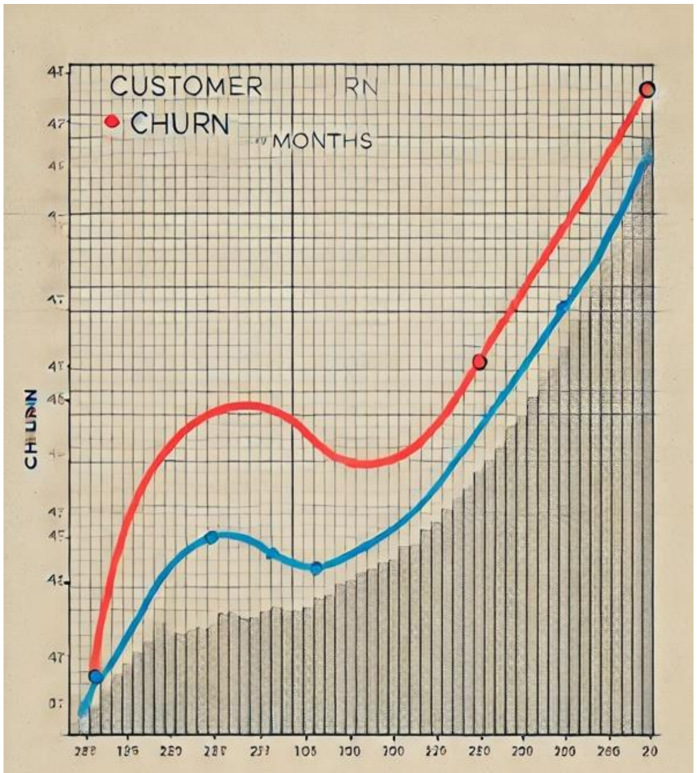
The main takeaway from this project is that machine learning models, particularly Random Forest, are effective tools for predicting customer churn. The Random Forest model was able to capture complex relationships between customer demographics, service usage, and billing information, making accurate predictions about customer churn. The application of the SMOTEENN technique played a crucial role in handling the class imbalance, allowing the model to perform well in identifying churned customers. While dimensionality reduction using PCA was explored, it did not improve the model’s performance in this case, suggesting that the feature set was already well-optimized.

The project highlights the potential of machine learning in real-world applications, where predictive models can provide valuable insights for businesses. By integrating this model into a live environment, businesses can monitor customer behavior in real-time and make data-driven decisions to reduce churn. The predictive power of the model can be used to identify at-risk customers early, allowing businesses to intervene with retention strategies such as personalized offers, improved customer service, or loyalty programs

### B.Future Works

While the model developed in this project achieved strong performance, there are several areas where future work could enhance the model's capabilities and extend its applicability:

- Incorporating Additional Features: Future models could benefit from incorporating additional data, such as customer interactions with support teams, social media engagement, and feedback from customer satisfaction surveys [5]. These features could provide a more comprehensive understanding of customer sentiment and behavior.
- Time-Series Analysis: One limitation of the current model is that it does not account for the temporal aspect of customer behavior. Future work could explore time-series models that track customer behavior over time, allowing for dynamic predictions that evolve as new data becomes available.
- Real-Time Deployment: The next step for this project is to deploy the model in a real-time environment. This could involve integrating the model into a customer relationship management (CRM) system or developing an API that allows businesses to make real-time predictions based on up-to-date customer data. The ability to make dynamic, real-time predictions would significantly enhance the practical value of the model.
- Improving Recall: While the model achieved a high accuracy, there is still room for improvement in recall. Future work could explore techniques such as cost-sensitive learning, where the model is penalized more heavily for misclassifying churned customers, to further reduce false negatives.



## REFERENCES

[1]. Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.( I)

[2]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.( II)

[3] Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer.( II)

[4] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.( III)

[5] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.( IV)