**CUSTOMER CHURN PREDICTION**

**USING DESCISION TREE AND RANDOM FOREST CLASSIFIER**

SRAVANI, LAHARI, DIVYA, SRUTHI

**ABSTRACTION:**

**The most critical challenge faced by the telecommunication banking and ecommerce is customer churn where customer retention directly impacts profit. This project aims to develop a ML model capable of predicting customer churn by analyzing the large dataset. The project has several stages: EDA, feature engineering, model building, model evaluation. Dataset was standardized to improve model performance. Here we have smoteenn technique(combining over-sampling and under-sampling approaches). Classification Algorithms like Logistic Regression, Random Forest, Support Vector Machine(SVM) are evaluated. Principal Component Analysis(PCA) was also applied for dimensionality reduction. The final model was saved using pickle library, making it accessible for deployment via an API.**

**This help companies to identify customers at risk of churn.**

**1**.**INTRODUCTION:**

Customer churn, also known as customer attrition, refers to the phenomenon where customers stop doing business with a company. For industries that rely heavily on customer subscriptions or repeat transactions, such as telecommunications, banking, and online services, churn can have a significant financial impact. It is often more expensive to acquire new customers than to retain existing ones, making churn prediction an essential focus for companies aiming to enhance customer retention. This project aims to develop a machine learning model that can accurately predict which customers are likely to churn, allowing businesses to take proactive measures to retain them.

Machine learning models have proven highly effective in predicting churn by analyzing vast amounts of data related to customer behavior, transaction history, and service usage. Unlike traditional statistical models, machine learning models can identify complex, non-linear relationships in the data, making them highly suitable for churn prediction tasks. By leveraging techniques such as Random Forest, Support Vector Machines, and data balancing techniques like SMOTEENN, this project provides a robust solution for churn prediction.

The dataset used in this project includes features related to customer demographics, service subscriptions, and account history. By preprocessing the data and applying various machine learning techniques, the goal is to develop a predictive model that can offer insights into customer behavior. The outcome of this project is expected to aid decision-makers in developing effective customer retention strategies, ultimately improving the company's bottom line.

**2.LITERATURE REVIEW:**

Customer churn prediction has been a subject of extensive research due to its critical importance in industries such as telecommunications, financial services, and retail. Traditional statistical models, such as Logistic Regression and Decision Trees, have been widely used to predict churn. However, the advent of machine learning has brought more advanced techniques into the spotlight, including Random Forest, Gradient Boosting, and Support Vector Machines, all of which offer better performance in handling complex datasets.

One key challenge in churn prediction is dealing with imbalanced datasets. In most cases, the number of customers who churn is significantly smaller than those who remain, leading to skewed data distributions. Researchers have proposed various techniques to address this imbalance, with SMOTE (Synthetic Minority Over-sampling Technique) being one of the most popular. SMOTE works by generating synthetic data points for the minority class to balance the dataset. However, SMOTE alone can lead to overfitting, which is why it is often combined with techniques like Edited Nearest Neighbor (ENN), forming the SMOTEENN method used in this project.Another challenge is dimensionality reduction, which involves reducing the number of features while preserving the most relevant information. Techniques like Principal Component Analysis (PCA) are commonly applied in churn prediction studies to simplify models and improve generalization. Although PCA can be useful in many cases, it may not always lead to performance improvements, as seen in this project. Overall, this study builds on existing literature by employing a combination of SMOTEENN for data balancing and Random Forest for model building, resulting in high prediction accuracy.

**3.METHODOLGY:**

3.1 Data Collection and Preprocessing

The dataset used in this project contains various features representing customer demographics (age, gender, tenure), service usage (internet service, phone service), and account-related details (monthly charges, total charges). The data was obtained from a telecom company's customer records, where the primary target variable is whether a customer churned (binary classification: churn or no churn). The dataset was split into training and test sets, ensuring a balanced representation of both churned and non-churned customers.

Before building any models, several preprocessing steps were applied. Categorical variables such as gender, internet service, and contract type were encoded into numerical values to make them suitable for machine learning algorithms. Additionally, feature scaling was performed on continuous variables like monthly charges and total charges to standardize the data and avoid skewing the results. Missing values were handled appropriately by either imputing them or removing rows with significant missing data.

3.2 Exploratory Data Analysis (EDA)

EDA was conducted to gain insights into the distribution of features and their relationship with the target variable (churn). For instance, customers with shorter tenure and higher monthly charges were found to have a higher likelihood of churn. Visualizations such as histograms, bar plots, and correlation matrices were created to explore relationships between features. Key observations from EDA informed the feature engineering process, such as grouping tenure into categories to simplify the model's complexity.

3.3 Model Selection and Training

Several machine learning models were evaluated for their effectiveness in predicting customer churn. Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machines (SVM) were implemented, each with different strengths in handling classification tasks. Cross-validation was used to fine-tune hyperparameters and prevent overfitting. Based on initial performance metrics such as accuracy, precision, and recall, the Random Forest classifier emerged as the best-performing model.

3.4 Addressing Data Imbalance

Given the significant class imbalance, with only a small percentage of customers labeled as churned, the SMOTEENN technique was applied. This method combines SMOTE, which synthetically generates data points for the minority class, and ENN, which removes noisy data points from the majority class. This combination helps balance the dataset while reducing the likelihood of overfitting, making it particularly effective for this project.

3.5 Dimensionality Reduction

Although the dataset included many features, not all of them contributed equally to the model’s performance. Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data, simplifying the model while maintaining its predictive power. However, PCA did not significantly improve performance, so the final model did not include PCA.

**4.RESULTS AND DISCUSSION:**

The Random Forest classifier with SMOTEENN produced the best results, achieving an accuracy of 94.27% on the test dataset. The model's high precision and recall indicated that it was able to effectively distinguish between churned and non-churned customers. Below are the detailed metrics for the final model:

|  |  |
| --- | --- |
| **METRIC** | **VALUE** |
| Accuracy | 94.27% |
| Precision | 0.72 |
| Recall | 0.81 |
| F1-Score | 0.77 |

The confusion matrix showed that the model was able to correctly predict a large number of both churned and non-churned customers. Precision and recall values suggest that the model is well-balanced and does not disproportionately favor one class over the other. However, there is still some room for improvement, particularly in reducing the number of false positives (customers predicted to churn but who do not).

A key takeaway from this project is the importance of addressing data imbalance. The application of SMOTEENN significantly improved model performance by ensuring that the model was not biased toward the majority class. This is a crucial consideration in churn prediction, where the minority class (churned

customers) is often the most important for business decision-making.Although PCA was applied for dimensionality reduction, it did not lead to performance gains,

suggesting that the original feature set was well-optimized. Future studies could explore additional feature selection methods or external data sources to further enhance the model's accuracy and generalizability.

**5.CONCLUSION AND FUTURE WORK:**

This project demonstrates the successful application of machine learning techniques for customer churn prediction. By using a combination of Random Forest classification and SMOTEENN for data balancing, the model achieved a high accuracy of 94.27%. The importance of addressing class imbalance cannot be overstated, as it had a significant impact on the model's performance. Additionally, the Random Forest model proved to be robust in handling a variety of feature types, including categorical and continuous data, making it an

ideal choice for churn prediction tasks.

In terms of future work, there are several directions in which this project could be expanded. First, additional features, such as customer support interactions or external market data, could be integrated into the model to improve its predictive power. Second, advanced ensemble techniques like XGBoost or LightGBM could be explored to see if they offer better performance compared to Random Forest. Lastly, deploying this model into a real-time environment and integrating it with a customer relationship management (CRM) system would allow businesses to make dynamic predictions and take immediate actions to retain at-risk customers.

**6.REFERENCES:**

* Brownlee, J. (2020). Imbalanced Classification with Python. Machine Learning Mastery.
* Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning. Springer.
* 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.
* Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
* Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Proceedings of the 2005 International Conference on Intelligent Computing.