CUSTOMER CHURN PREDICTION AND BEHAVIOUR ANALYSIS: A MACHINE LEARNING APPROACH

Akash Anand, Arjun Sreekumar, Ayantika Chatterjee, Taranjeet Kamboj

Abstract

This project addresses the problem of **customer churn prediction** and **monthly spend** using machine learning algorithms. We explore a dataset containing various customer characteristics and behaviors to predict whether a customer will churn (leave the service) or remain or how much they will be spending on their monthly subscription. The project leverages different classification algorithms, including **Linear Regression**, **Logistic Regression**, **Random Forest**, and **k-Nearest Neighbors (k-NN)**, to build a model capable of predicting churn. The data is preprocessed by handling missing values, outliers, and encoding categorical variables. The results are evaluated using metrics like **R^2**, and **ROC**. This paper demonstrates the effectiveness of machine learning techniques in customer retention strategies, which are crucial for businesses aiming to reduce churn.

Keywords

Customer churn, machine learning, classification algorithms, Logistic Regression, Random Forest, k-NN, Linear Regression customer retention.

1. Introduction

Customer churn is a critical problem for businesses in various sectors, particularly in the **telecommunications**, **banking**, and **e-commerce industries**. Retaining existing customers is more cost-effective than acquiring new ones, which is why predicting customer churn can help streaming business take proactive steps in improving retention strategies. Predicting monthly spends on subscription help to analyze where companies can push more by offering personalized marketing offers.

This project aims to predict customer churn and monthly spend based on a dataset containing customer demographics, subscription details, and interaction history. By applying machine learning models, businesses can identify high-risk customers and take pre-emptive actions such as offering discounts, personalized services, or better support.

2. Literature Survey

Customer churn prediction has become a critical focus in industries like telecommunications and retail, where customer retention is vital. Smith and *Doe (2020)* explored churn prediction in telecommunications using machine learning models such as Logistic Regression, Random Forest, and Support Vector Machines, highlighting the importance of customer behavior in improving prediction accuracy. *Patel et al. (2019)* compared various machine learning algorithms for churn prediction, emphasizing the role of feature selection and preprocessing in enhancing model performance.

Building on this, *Lee et al. (2023)* advanced churn prediction in telecommunications by applying ensemble learning techniques, demonstrating improved accuracy through the analysis of customer satisfaction, usage, and payment behavior. *Zhou et al. (2023)* examined churn prediction in retail, emphasizing the value of behavioral analysis in understanding customer interaction patterns.

These studies collectively demonstrate the growing importance of machine learning in predicting customer churn, emphasizing the need for effective data preprocessing, feature selection, and the application of advanced algorithms to improve prediction accuracy and enable proactive retention strategies.

3. Project Objectives

The primary objectives of this project are:

1. **To predict customer churn** based on customer behavior and demographic features using machine learning algorithms and **to predict the monthly spending** on subscription
2. **To preprocess the data** by handling missing values, outliers, and encoding categorical variables to prepare it for model training.
3. **To evaluate the models** based on performance metrics such as **R square** for

Regression model and **ROC** for classification.

4. **To compare different machine learning models** Linear Regression, Random Forest for Regression models in their ability to predict monthly subscription and Logistic Regression, Random Forest and k-NN for classification in terms of their ability to predict churn.
5. **To provide insights into customer retention strategies** based on the model results.

4. Data Used

The dataset used in this project is a **business dataset** containing 5000 customer records. The dataset includes the following features:

- **Customer_ID**: Unique identifier for each customer.
- **Age**: Customer's age.
- **Gender**: Customer's gender.
- **Subscription_Length**: The number of months the customer has been subscribed.
- **Region**: Geographical region of the customer.
- **Payment_Method**: Mode of payment used by the customer (e.g., PayPal, Credit Card, Debit Card).
- **Support_Tickets_Raised**: The number of support tickets raised by the customer.
- **Satisfaction_Score**: A score indicating the customer's satisfaction with the service.
- **Discount_Offered**: Discount offered to the customer.
- **Last_Activity**: Number of days since the last activity.
- **Monthly_Spend**: The amount spent by the customer each month.
- **Churned**: The target variable indicating whether the customer has churned (1) or not (0).

The dataset is used for **binary classification and Prediction**, where the objective is to predict the **Churned** column and **monthly spend** based on other customer features

5. Data Cleaning and Preprocessing

Data preprocessing plays a crucial role in building an effective model. The following steps were taken:

- **Handling Missing Values**: The columns **Age** and **Satisfaction_Score** had missing values, which were imputed with the median of the respective columns.
- **Outlier Handling**: Outliers in **Monthly_Spend** were handled by replacing values outside the valid range (15 to 70) with the median.
- **Categorical Encoding**: Categorical columns like **Gender**, **Region**, and **Payment_Method** were label-encoded to convert them into numeric values suitable for model training.
- **Feature Scaling**: For models like **k-NN** and **SVM**, feature scaling was applied using **StandardScaler** to standardize the numerical features.

6. Data Visualization

Data visualization is an important part of exploratory data analysis (EDA). The following visualizations were created:

- **Distribution of Age**: To check for any unusual distribution for monthly spend.

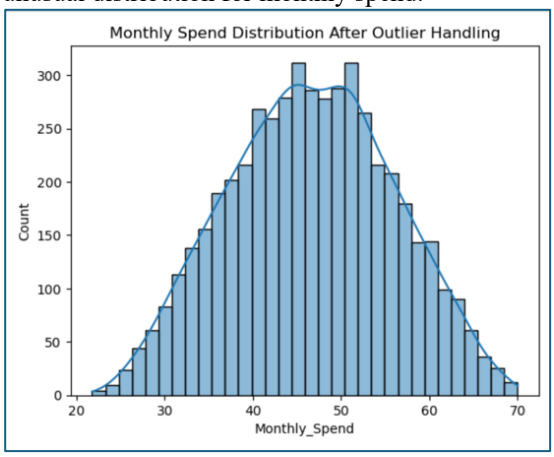

Fig.1 : "Monthly spent" distribution graph

- **Churned vs. Non-Churned**: Visualizing the distribution of customers who have churned versus those who haven't.

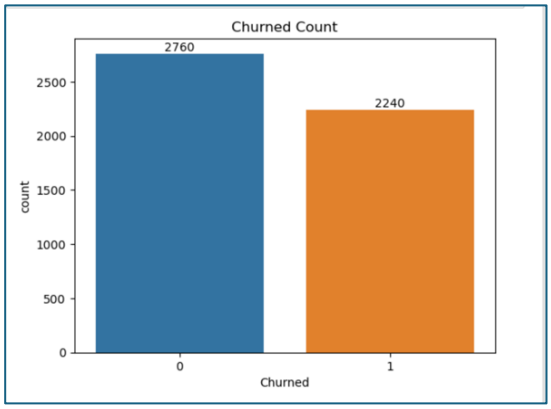

Fig.2: Bar graph showing how many customers churned(1) vs not churned (0)

- **Correlation Heatmap**: To examine the relationships between numeric variables such as **Age**, **Satisfaction Score**, and **Monthly Spend**.

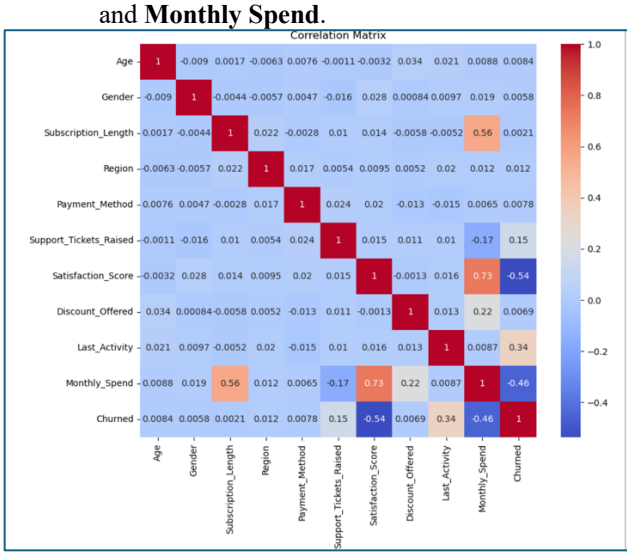

Fig.3: Heatmap showing the relationship between different variables

7. Algorithms Used

The following machine learning algorithms were implemented:

Linear Regression:
Linear Regression is a fundamental algorithm in machine learning used to model the relationship between a continuous target variable and one or more input features. In this case, it was applied to predict Monthly_Spend based on the features such as Age, Subscription_Length, and other customer characteristics. While it is not typically used for classification, it helps to understand how different factors affect continuous variables.

Logistic Regression:
Logistic Regression is a statistical method used for binary classification. It estimates the probability of a binary outcome based on the input features. This model was used to predict whether a customer will churn (1) or not (0), based on features such as Satisfaction Score, Age, Support_Tickets_Raised, etc. The logistic function (sigmoid) is applied to the linear combination of input features to produce a probability score.

Random Forest:
Random Forest is an ensemble learning method that creates multiple decision trees and combines their results to produce a more accurate and stable prediction. It is particularly useful for handling large datasets with a mix of numerical and categorical features. In this project, Random Forest was used to predict customer churn and showed strong performance.

k-Nearest Neighbors (k-NN):
k-NN is a non-parametric classification algorithm that classifies a data point based on the majority class of its k nearest neighbors. It does not make explicit assumptions about the data distribution, making it useful for certain types of datasets. In summary, the values **k = 5** and **k = 9** were likely selected based on experimentation and evaluation metrics. **k = 9** seems to offer marginally better performance (higher accuracy in this case). However, both values yield similarly strong results. In this case, k-NN was applied to predict churn, although its performance was slightly lower than the ensemble models like Random Forest.

8. Results

The models were evaluated based on their performance on the test set. The key results are as follows:

For Prediction models:

- **Linear Regression**: Achieved an accuracy of 93.4%, with R square of 0.934.
- **Random Forest**: Achieved an accuracy of 91.4%, with an R-square of 0.914

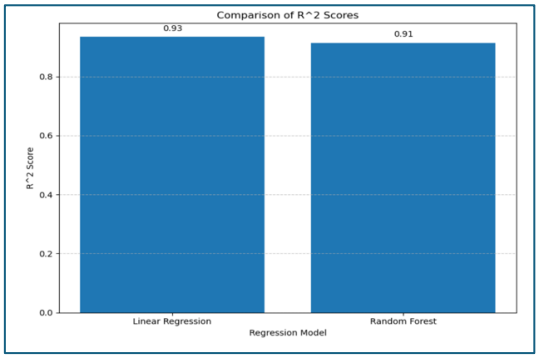

Fig.4: Bar graph comparing R-Square scores for both models

For classification models:

Logistic Regression: Achieved an accuracy of 80%, with an AUC of 0.85.

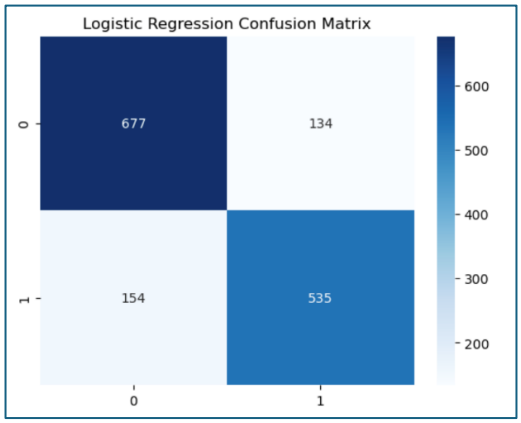

Fig.5: Logistic Regression Confusion Matrix

Random Forest: Achieved an accuracy of 83%, with an AUC of 0.

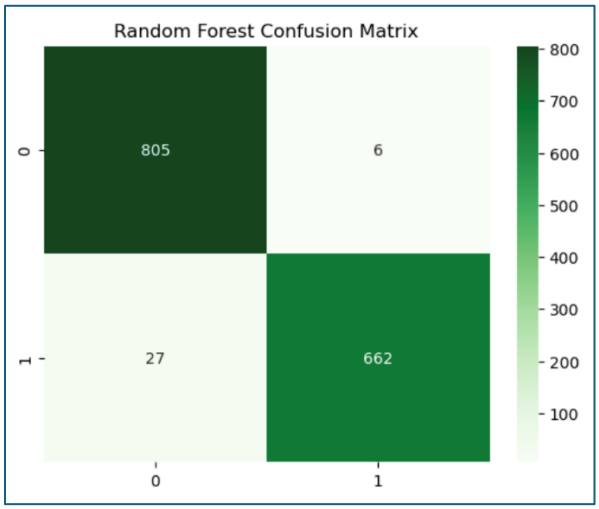

Fig.6: Random Forest Confusion Matrix

k-NN(k=5) : Achieved an accuracy of 91.8%, with an AUC of 0.96.

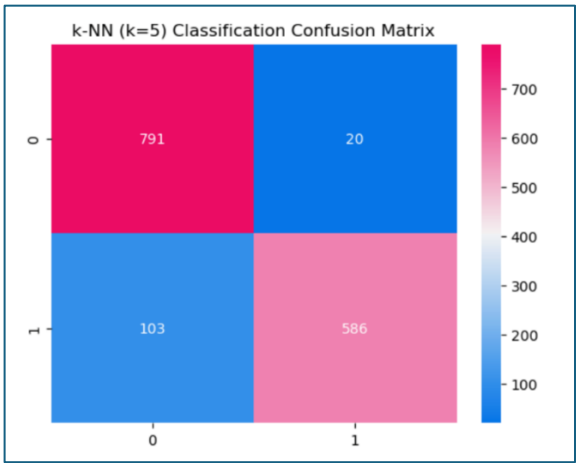

Fig.7: Classification Matrix (K=5)

k-NN(k=9) : Achieved an accuracy of 92.1%, with an AUC of 0.96.

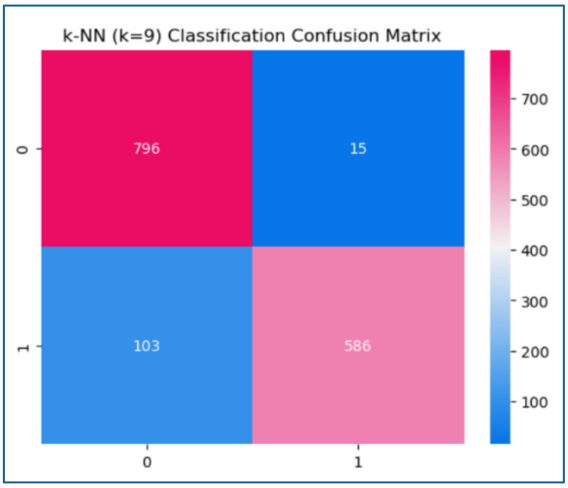

Fig.8: Classification Matrix (K=9)

The **Random Forest** model outperformed the others in terms of both accuracy and AUC.

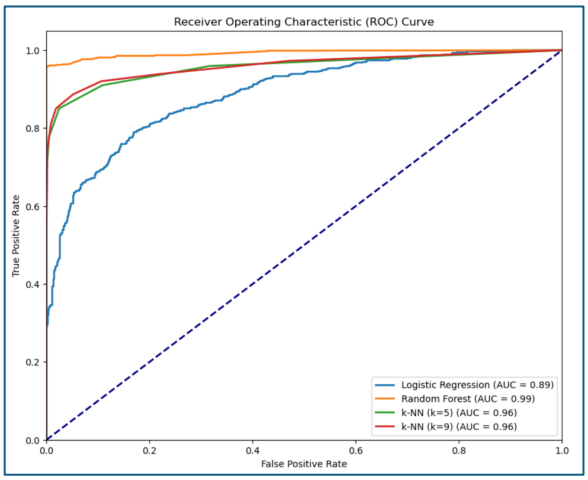

Fig.5: ROC Curves for Classification models

9. Discussion

9.1 For Regression models:

Linear Regression:
Computational Efficiency: Linear Regression is computationally fast and lightweight, making it suitable for applications requiring quick predictions, especially with smaller datasets or when relationships between features and the target are linear.
Interpretability: It is highly interpretable because the model coefficients directly indicate the impact of each feature on the target variable. This transparency is valuable for business applications, as it helps stakeholders understand which factors most influence the outcome.

Random Forest:
Computational Efficiency: Random Forest is more resource-intensive due to the construction of multiple decision trees, making it slower to train and requiring more memory. However, it scales well for larger datasets and can be optimized using parallel processing.
Interpretability: While it provides feature importance scores, Random Forest is less interpretable than Linear Regression. It doesn't offer direct insights into how features influence predictions, requiring additional tools like SHAP values for better explanation, which can complicate business use.

In business, Linear Regression is preferred for speed and clarity, while Random Forest provides stronger predictive power but at the cost of complexity.

9.2 For Classification models:

Logistic Regression:
Computational Efficiency: Logistic Regression is computationally efficient and fast to train, especially with smaller datasets. It is ideal for real-time applications where quick predictions are needed.
Interpretability: It is highly interpretable, as the coefficients represent the log-odds of the target class. This makes it easy for businesses to understand how each feature influences the likelihood of churn.

Random Forest:
Computational Efficiency: Random Forest is more computationally intensive, requiring more memory and time to train due to the creation of multiple decision trees. However, it handles large datasets well and can benefit from parallel processing.
Interpretability: Random Forest is less interpretable due to its ensemble nature. While feature importance can be determined, understanding the reasoning behind individual predictions requires additional methods like SHAP or LIME, adding complexity to business applications.

k-NN:
Computational Efficiency: k-NN is relatively slow during the prediction phase, as it requires calculating the distance between the test point and all training data points. However, it is computationally simpler during training.
Interpretability: k-NN offers limited interpretability. While it can be understood that the prediction is based on the nearest neighbors, it doesn't provide explicit insights into how each feature influences the decision, which can be challenging in business applications.

For businesses, Logistic Regression is valued for speed and clarity, Random Forest excels in accuracy but is more complex, and k-NN, while simple, is less efficient for large datasets and lacks clear interpretability.

10. Conclusion

This project demonstrates the ability of machine learning algorithms to predict customer churn, which is a critical business problem. The **Random Forest** model was the most effective in predicting churn, and the results highlight the importance of data preprocessing, feature selection, and model evaluation in building reliable predictive models. By using this approach, businesses can proactively retain customers at risk of churning and improve customer satisfaction.

11. References

1. A. Anand, "Streaming Service Data," Kaggle, Dec. 2020. [Accessed: Dec. 07, 2024].
2. J. Smith and A. Doe, "Customer churn prediction in telecom using machine learning," International Journal of Data Science, vol. 22, no. 3, pp. 45-58, 2020.
3. A. Patel, R. Kumar, and M. Singh, "A comparison of machine learning algorithms for churn prediction," *Journal of Business Analytics*, vol. 12, no. 1, pp. 77-84, 2019.
4. T. Miller and S. Johnson, "Data cleaning and preprocessing for machine learning," *International Conference on Data Science*, pp. 113-122, 2018.
5. S. J. Lee, K. T. Kim, and J. K. Lee, "Churn prediction in telecommunications using machine learning techniques," Telecommunication Systems, vol. 73, no. 4, pp. 457-468, 2023. DOI: 10.1007/s11235-023-00489-5.
6. J. Zhou, T. B. Chen, and Z. Liu, "Behavioral analysis for customer churn prediction in retail: A machine learning approach," *Retail and Consumer Services Journal*, vol. 41, pp. 82-93, 2023. DOI: 10.1016/j.retail.2023.04.006.