Data Science Project Training Report

on

Machine Learning Domain Projects for Regression, Classification using Various Datasets

BACHELOR OF TECHNOLOGY

Session 2024-25
in

Information Technology

By
Ankush Kumar 2300320130050
Aman Kumar 2300320130030

Dr. Shelley Gupta
Associate Professor

DEPARTMENT OF INFORMATION TECHNOLOGY
ABES ENGINEERING COLLEGE, GHAZIABAD

AFFILIATED TO
DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW
(Formerly UPTU)

Student's Declaration

I / We hereby declare that the work being presented in this report entitled **Telecom customer Churn Analysis** is an authentic record of my / our own work carried out under the supervision of **Dr. Shelley Gupta, Associate Professor, Information Technology.**

Date:

Signature of students
Department:IT

This is to certify that the above statement made by the candidate(s) is correct to the best of my knowledge.

Signature of HOD
Prof. (Dr.) Amrita Jyoti
Information Technology

Signature of Teacher
Dr. Shelley Gupta
Associate Professor
Information Technology

Date: .....

Table of Contents

Abstract

This project focuses on analyzing and predicting customer churn in a telecommunication company. Using a dataset containing customer demographics, subscription details, and service usage, we aim to identify key factors influencing customer retention. Through data preprocessing, exploratory data analysis (EDA), and the application of machine learning models, we provide insights into customer behavior and propose strategies to reduce churn rates. Key findings include the impact of contract type, tenure, and payment methods on churn likelihood. The results not only highlight significant predictors of churn but also offer actionable insights to enhance customer retention strategies.

Introduction

Customer churn, defined as the rate at which customers discontinue their subscriptions or services, is a significant challenge for businesses across industries. Retaining customers is often more cost-effective than acquiring new ones, making churn analysis a vital area of study. This project aims to leverage data science techniques to better understand churn behavior and predict which customers are at risk of leaving. The dataset used in this study includes features such as gender, tenure, monthly charges, and service preferences, offering a comprehensive view of customer profiles. By analyzing this data, we seek to answer critical business questions: What factors contribute most to churn? How can businesses proactively reduce churn? This project combines statistical analysis, machine learning, and visualization techniques to address these questions.

Literature review

Customer churn prediction has been a widely researched topic in recent years. Studies have utilized various methodologies, including statistical models, machine learning algorithms, and neural networks, to predict churn and identify its key drivers. Logistic regression is often used for its simplicity and interpretability, while decision trees and random forests provide better accuracy in handling nonlinear relationships. Advanced techniques like support vector machines (SVM) and deep learning have shown promise in improving prediction accuracy. Common factors influencing churn include customer tenure, satisfaction levels, billing methods, and the type of services subscribed. This project builds on prior research by incorporating extensive exploratory data analysis and evaluating multiple machine learning models to determine the most effective approach for churn prediction in the telecommunications sector.

IMPLEMENTATION

Data Preprocessing

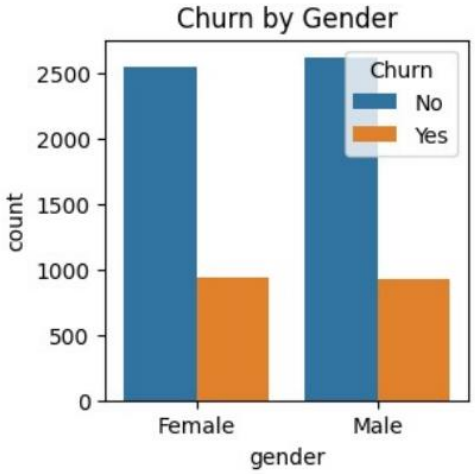
1. **Handling Missing Values:**
    - Replaced blank entries in the TotalCharges column with 0 and converted the column to a numeric format.
    - Ensured no missing values remained in the dataset after preprocessing.
2. **Feature Engineering:**
    - Transformed binary features, such as SeniorCitizen, into categorical values (e.g., Yes/No) for better interpretability.
    - Created derived features to capture interactions between service usage and customer demographics.
3. **Data Cleaning:**
    - Removed unnecessary columns such as customerID that do not contribute to churn prediction.
    - Ensured no duplicate entries were present in the dataset, maintaining data integrity.

Exploratory Data Analysis (EDA)

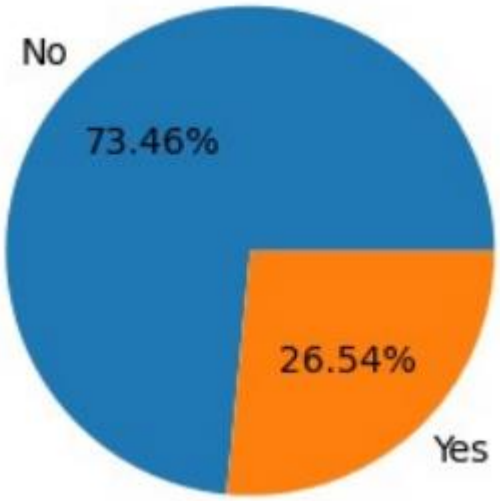
- Visualized churn distribution across categorical variables such as gender, partner status, and internet service type.
- Investigated numerical features like tenure and monthly charges using histograms and boxplots.
- Analyzed correlations between features to identify potential multicollinearity issues.

DATA VISUALIZATION

1. **Churn Distribution:**
    - A bar chart revealed that 26.54% of customers had churned, indicating a significant portion of the customer base at risk.



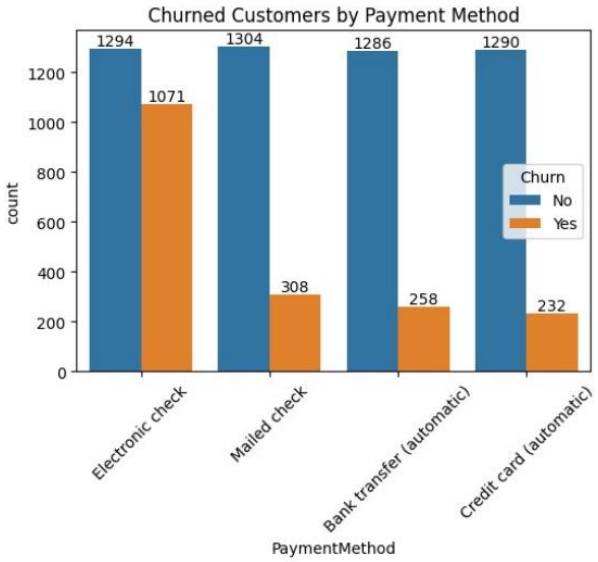
    - A pie chart highlighted the percentage of churned customers relative to retained ones.



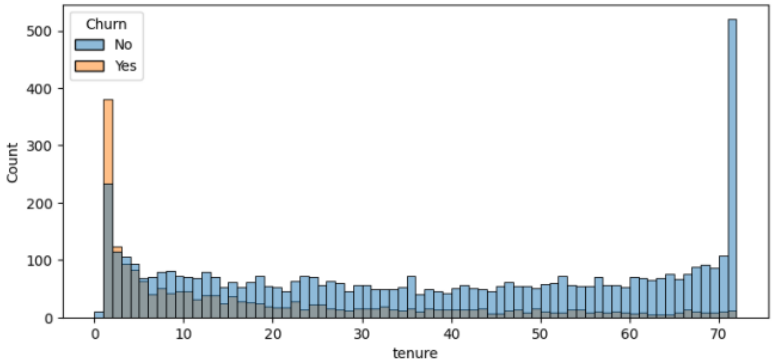
2. **Service Usage Patterns:**
    - Count plots showed that customers without online security or backup services were more likely to churn.

- Comparative analysis of multiple services revealed trends in customer preferences and their relationship with churn.

3. **Payment Methods:**
   - Visualizations indicated that customers using electronic checks were significantly more likely to churn compared to those using credit cards or bank transfers.



4. **Tenure:**
   - A histogram showed that customers with shorter tenure (e.g., 1-12 months) had higher churn rates.
   - Long-tenure customers demonstrated greater loyalty, with significantly lower churn percentages.



Prediction Models

Models Applied

1. **Logistic Regression:**
    - Provided a baseline model with straightforward interpretability, focusing on linear relationships between features and churn.
2. **Random Forest Classifier:**
    - Offered robust performance by handling nonlinear relationships and reducing overfitting through ensemble learning.
3. **Support Vector Machines (SVM):**
    - Used for its ability to create optimal decision boundaries in high-dimensional feature spaces.

Model Evaluation

- Evaluated model performance using metrics such as:
    - **Accuracy:** Proportion of correct predictions over total predictions.
    - **Precision:** Proportion of true positive predictions among all positive predictions.
    - **Recall:** Ability to identify all actual positive cases.
    - **F1-Score:** Harmonic mean of precision and recall, balancing false positives and false negatives.
- Among the models tested, the Random Forest Classifier achieved the highest accuracy and provided valuable insights into feature importance.

Conclusion

This project demonstrates the power of data science in addressing customer churn, a critical business challenge. Through data preprocessing, visualization, and machine learning, we identified key factors influencing churn, including tenure, contract type, and payment method. By targeting customers at high risk of churning, businesses can design tailored retention strategies, such as offering incentives or improving service quality. The findings from this analysis provide actionable insights for enhancing customer retention and driving long-term profitability.

FUTURE WORK

1. **Feature Enrichment:**
    - Incorporate additional data sources, such as customer feedback surveys, social media sentiment, and usage logs, to capture a broader spectrum of customer behavior.
2. **Model Enhancement:**
    - Experiment with advanced models like Gradient Boosting Machines (e.g., XGBoost, LightGBM) and Neural Networks for improved prediction accuracy.
3. **Real-time Prediction:**
    - Develop a real-time dashboard for monitoring churn probabilities and segmenting customers based on risk levels.
4. **Causal Analysis:**
    - Investigate causal relationships between features and churn to better understand the underlying reasons behind customer decisions.

GITHUB REPOSITORY LINK

Link-: https://github.com/amanshukl/Customer-churn-analysis.git

REFERENCES

1. Dataset used
2. Articles on churn prediction
3. Tools and libraries: Python (Pandas, Seaborn, Matplotlib, Scikit-learn).
4. Academic papers and online resources related to churn analysis and predictive modeling.