D-EKALE / phase-3-final-project Public

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main Go to file Add file Code

D-EKALE presentation ... 1 minute ago 11

README.md	'Final Updates'	3 hours ago
phase-3-final-project.ipynb	Final Notebook	1 hour ago
presentation.pdf	presentation	1 minute ago
syria-tel.csv	Made some updates	11 hours ago

README.md

Predicting Customer Churn in the Telecommunications Industry

A Data-Driven Approach to Improve Customer Retention

image.png

Project Overview:

This project aims to develop a classification model that will predict customer churn for SyriaTel, a telecommunications company. I have chosen to follow the CRISP-DM method to complete this project. It will involve six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The project purposes to provide insights into the patterns and factors influencing customer churn, and also develop a predictive model to assist in reducing customer attrition.

Business Understanding:

SyriaTel is the major stakeholder for this project. They are interested in reducing customer churn. By helping them predict customer churn, they can take proactive measures to ensure maximum customer retentions and profit maximization. The project majorly focuses on identifying patterns that facilitate to customer churn and providing recommendations on how to mitigate this.

Data Understanding:

We are anlysing the SyriaTel.csv dataset that is available on Kaggle. The dataset has 3333 rows and 21 columns and has no null values or duplicates. Therefore we don not need to impute any missing values or drop any duplicated values in this case. Among the 21 columns five of them are categorical in nature; `state` , `phone number` , `international plan` , `voice mail plan` , `churn` . `Churn` which is our target variable in the data set is of boolean data type. Thus, we will make it binary later when building our models.

Some of the columns based on domain knowledge are not actually good predictors and thus dropping them before fitiing into our models will be good. For example, the `phone number` variable has nothing to do with customer chruning the company.

Most values in the dataset are numerical in nature. The summary statistics provides a brief overview of the dataset and the range of values observed in each numerical column.

Data Preparation

To prepare our data for modeling, we will:

- Remove outliers from the dataset
- Address multicollinearity by removing highly correlated features

About

No description, website, or topics provided.

Readme

0 stars

2 watching

0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● Jupyter Notebook 100.0%

- Perform scaling on the numerical variables
- Handle class imbalance using the SMOTE method
- Proceed with building models for predictions.

Modeling

Model Building: Three different models, including Logistic Regression, Random Forest, and Support Vector Machine (SVM), are trained and evaluated using appropriate evaluation metrics such as accuracy, precision, recall, F1 score, and ROC AUC score.

Model Comparison: The performance of the models is compared based on evaluation metrics and ROC curves to identify the best-performing model.

Requirements

Python 3.x Required libraries: pandas, numpy, scikit-learn, matplotlib, seaborn

Usage

Clone the repository: git clone https://github.com/D-EKALE/phase-3-final-project Install the required libraries: pip install pandas numpy scikit-learn matplotlib seaborn Run the data preprocessing steps to handle missing values, duplicates, and outliers. Perform exploratory data analysis to gain insights into the dataset. Build and train different models (Logistic Regression, Random Forest, SVM) using appropriate evaluation metrics. Compare the performance of the models and select the best-performing model.

Contributors

Daniel Ekale

License