charlesot / DSF-PT08P3-Phase-3-project-

Code Pull requests Actions Projects Wiki Security Insights Settings

DSF-PT08P3-Phase-3-project- Public

forked from Charlesotwori/DSF-PT08P3-Phase-3-project-

1 Branch 0 Tags Go to file t Go to file Add file Code

This branch is 26 commits ahead of Charlesotwori/DSF-PT08P3-Phase-3-project-:main .

Contribute Sync fork

charlesot updated ipynb file	1a31380 · 1 hour ago	
.ipynb_checkpoints	update	3 hours ago
data	Add index file	4 days ago
Readme.md	Update Readme.md	10 hours ago
index.ipynb	updated ipynb file	1 hour ago

README

DSF-PT08P3- Phase 3 project

About

DSF-PT08P3- Phase 3 project

Readme
Activity
0 stars
0 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%

SyriaTel Project: Reducing Customer Churn Through Predictive Analysis

1. Project Overview

SyriaTel is a telecommunication company that is interested in reducing the number of customer churn/attrition to reduce loss of revenue. The project will enable the reduction of churn through analytics of historical data and the development of machine learning models to predict customers who are likely to churn and develop a mitigation strategy. The analysis involves data understanding, data cleaning, preprocessing, building predictive models, evaluating their performance, and visualizing key insights. The logistic Regression model is used as a base model. Other models used are Decision tree, XGBoost, and Random forest.

2. Business and Data Understanding

Telecom companies experience customer churn/attrition when customers voluntarily cease their relationship with the company. This leads to financial loss due to revenue reduction and market share loss. Acquiring a new customer is more expensive than maintaining existing ones. During churn, the company also loses the future revenue from the customer and the resources spent to acquire the customer. Thus, reducing profitability. Beyond financial loss, high customer churn can indicate a deeper problem with the company regarding customer service, product appeal, or quality of processes. This can erode the company's reputation and further erode the market share.

- Analytics of churn will provide insights into why customers churn and identify customers who are likely to leave so that a targeted strategy can be developed

to convince them to stay.
- The Stakeholder audience for the project is SyriaTel telecommunication company executives from the marketing, sales, and innovations departments.

Main Objective

- Predict which customers are likely to churn

Secondary objectives

- preprocess churn data
- Train and evaluate multiple classification models
- Visualise the evaluations and feature importance
- Recommend a model to predict customers who are likely to churn and other insights to reduce churn

The predictors (features)

Include the following:-

- account length
- international plan
- voice mail plan
- Number of voicemail messages
- total day minutes used
- day calls made
- total day charge
- total evening minutes
- total evening calls
- total evening charge

- total night minutes
- total night calls
- total night charge
- total international minutes used
- total international calls made
- total international charge
- Number of customer service calls made

Target Variable:

Churn; If the customer has churned (1 = yes; 0 = no)

How analysis is run

prerequisites

- Python 3.8 +

Import the following libraries

- import pandas as pd
- import seaborn as sns
- import matplotlib.pyplot as plt
- import numpy as np
- from sklearn import preprocessing
- from sklearn.preprocessing import OneHotEncoder
- from sklearn.preprocessing import StandardScaler
- from imblearn.over_sampling import SMOTE

- from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score
- from sklearn.linear_model import LogisticRegression
- from sklearn.tree import DecisionTreeClassifier, plot_tree
- from sklearn.ensemble import RandomForestClassifier
- from xgboost import XGBClassifier
- from sklearn.metrics import roc_auc_score, roc_curve, precision_recall_curve, accuracy_score, confusion_matrix, classification_report

load the data set or clone it (https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset)

Exploratory data analysis

- Checked for missing values and duplications
- Visualised distributions, boxplots(Outliers and correlations to understand the features

3. Data Preparation

Preprocessing

- Dropped irrelevant column (phone number)
- one hot encoded categorical features and label encoded the target
- Standardized the numerical features
- Split the data into training (80%) and testing (20%)
- Addressed class imbalance by using SMOTE(Synthetic Minority oversampling Technique)

4. Modelling

- Four models were trained and evaluated

4.1 Logistic regression

- Baseline model
- Trained using LogisticRegression with class balancing
- Evaluated using accuracy, confusion matrix, ROC-AUC, precision-recall curve, and Classification report.
- Performance
  - Accuracy = 0.79
  - ROC-AUC = 0.83

4.2 Decision Tree

- Trained using the DecisionTreeClassifier no class balancing
- Evaluated using accuracy, confusion matrix, ROC-AUC, precision-recall curve, Classification report, and Feature importance.
- Performance
  - Accuracy 0.94
  - ROC_AUC = 0.83
  - Key features : total day minutes , customer service calls , total international charges

4.3 XGBoost

- Trained with XGBClassifier with class balancing
- Evaluated using accuracy, confusion matrix, ROC-AUC, precision-recall curve, Classification report, and Feature importance.

- Performance
  - Accuracy = 0.96
  - ROC-AUC = 0.89
  - Key features: Total day charge, total evening charge and customer service calls

4.4 Random Forest

- Trained using RandomForestClassifier with class balancing
- Evaluated using accuracy, confusion matrix, ROC-AUC, precision-recall curve, Classification report, and Feature importance.
- Performance
  - Accuracy = 0.95
  - ROC-AUC = 0.86
  - Key features: total day charges, total evening charges, and customer service calls

5. Evaluation

- Metrics used ;
  - Accuracy for overall correctness of prediction
  - confusion metric for a detailed breakdown of predictions
  - Precision, Recall, and F1 score on churn detection
  - ROC-AUC for model discrimination ability

6. Conclusion

- The XGBoost model had the overall best performance
- High feature importance in

- Customer service calls: Need to focus on the quality of customer service
- Usage metrics evening, day charges: need to consider incentives

7 Future work