Mosota-Kemunto-Gladys-2020 /
Gladys_phase_3_project

Code Issues Pull requests Actions Projects Wiki Security Insights Settinc

0 stars 0 forks 1 watching 1 Branch 0 Tags Activity

Public repository

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

Mosota-Kemunto-Gladys-2020 Remove mistakenly added file: h origin master

d34850e · 1 minute ago

Data	Add project files and presentations ...	2 hours ago
Images	Add images related to SyriaTel chur...	2 days ago
README.md	Update README.md: Added detaile...	2 hours ago
SyriaTel Churn Analysis Prese...	Final pdf presentation	7 minutes ago
SyriaTel_Churn_Powerpoint P...	Final PowerPoint presentation	22 minutes ago
Syrial Tel_ Churn_Poerpoint P...	Add new Word version of the Powe...	20 minutes ago
git.pdf	Add reference document: git.pdf	5 minutes ago
student.html	Add student.html: Exported notebo...	47 minutes ago
student_final.ipynb	Add student_final.ipynb: Final versio...	45 minutes ago
student_final.pdf	Finalized PDF version of the notebo...	45 minutes ago

README

Phase_3 Project - Data Science

Name: Gladys Kemunto
GitHub Repo: https://github.com/Mosota-Kemunto-Gladys-2020/Gladys_phase_3_project.git

Project Overview

Objective

SyriaTel, a leading telecommunications company in Syria, is facing substantial financial losses due to customer churn, which refers to the rate at which customers discontinue their relationship with a company within a specific period. Previously, SyriaTel focused on descriptive and inferential analyses to understand customer behavior and relationships between variables. To address the current churn issue, SyriaTel is shifting towards a predictive approach. The goal of this project is to develop a predictive model using the SyriaTel Customer Churn dataset that accurately identifies customers at risk of churning. By understanding which variables most influence churn, SyriaTel aims to implement targeted retention strategies, reduce churn rates, and improve customer loyalty and profitability.

Business Understanding

Customer churn is a critical challenge in the telecommunications industry that directly impacts profitability. SyriaTel's loss of customers due to churn prompted the need for advanced analytical techniques. Previously, the company focused on understanding the distribution of key variables and customer behavior, but this is no longer sufficient. To effectively combat churn, SyriaTel is adopting a predictive approach by building a model to forecast which customers are likely to churn. This predictive model will analyze features like customer demographics, usage patterns, and service interactions to determine their impact on churn.

Key Questions to Be Addressed

- **What is the best model for predicting customer churn?**
  After comparing various models, including Decision Tree, K-Nearest Neighbors (KNN), and Random Forest, the analysis will recommend the best overall performer based on metrics like accuracy, precision, recall, and ROC-AUC score.

- **How accurately can the model predict customer churn?**
  The analysis will evaluate the performance of various models using metrics such as accuracy, precision, recall, and the ROC-AUC score to determine how well they predict customer churn.

- **Which features are most influential in predicting customer churn?**
  Identifying the most impactful features, such as customer service interactions, usage patterns, and plan types, will help SyriaTel prioritize its retention efforts and design more effective interventions.

With these insights, SyriaTel can proactively identify at-risk customers and intervene with targeted strategies, reducing churn rates and improving customer satisfaction and loyalty.

Methodology for Machine Learning

To build a predictive model for identifying customers at risk of churning, the following steps were followed:

1. **Data Understanding:**
   Explore the dataset to understand its structure, feature types, and quality, checking for missing values and inconsistencies.

2. **Data Cleaning:**
   Prepare the data by handling missing values, correcting errors, and removing duplicates and irrelevant features.

3. **Exploratory Data Analysis (EDA):**
   Visualize data distributions and relationships between features and the target variable to gain insights and identify patterns.

4. **Data Preprocessing:**
   Transform the data for modeling by encoding categorical variables, scaling numerical features, and splitting into training and test sets.

5. **Modeling:**
   Train various models (e.g., Logistic Regression, Decision Trees, Random Forest) and use cross-validation to evaluate performance.

6. **Hyperparameter Selection:**
   Optimize model performance by tuning hyperparameters using techniques like Grid Search.

7. **Model Evaluation:**
   Assess models using metrics like accuracy, F1-score, and ROC-AUC; analyze confusion matrices and use tools like SHAP for interpretation.

8. **Recommendations and Conclusion:**
   Summarize findings, recommend strategies to reduce churn, and provide insights for future improvements.

Data Understanding

The dataset used is the **SyriaTel Customer Churn** dataset, containing 3333 rows and 21 columns with a mix of categorical, numerical, and boolean data types. It includes features such as state, account length, area code, phone number, international plan, voice mail plan, and various usage metrics like total day minutes and customer service calls. The target variable is **churn**, indicating whether a customer has churned (True) or not (False).

Data Cleaning

The data cleaning process involved handling missing values, duplicates, and irrelevant features (e.g., phone number). Categorical variables were encoded using One-Hot Encoding, and numerical features were scaled. Features with high correlation, such as total charges and minutes, were identified to reduce redundancy.

Exploratory Data Analysis (EDA)

Key observations included:

- Potential outliers in features like voicemail messages and customer service calls.
- Distribution analysis showed class imbalance in churn, necessitating handling techniques like SMOTE.
- Correlation analysis identified redundancy in highly correlated features, which were removed to improve model efficiency.

Preprocessing

The data was split into training and test sets, categorical variables were encoded, and features were scaled. Class imbalance was addressed using SMOTE, resulting in a balanced dataset for training.

Modeling and Evaluation

Four models were evaluated:

1. **K-Nearest Neighbors (KNN):**
   Showed moderate performance but struggled with test set generalization.

2. **Logistic Regression:**
   Demonstrated consistency between training and test data but had modest precision.

3. **Decision Tree:**
   Performed well with balanced accuracy, precision, and recall.

4. **Random Forest:**
   Emerged as the top performer with high accuracy and robust generalization.

Fine-Tuning and SHAP Analysis

Fine-tuning focused on the Random Forest and Decision Tree models. SHAP analysis identified the most influential features affecting churn predictions, such as international plan usage and customer service calls.

Key Insights and Recommendations

1. **Deploy the Random Forest Model:**
   With the best overall performance, the Random Forest model is recommended for deployment to accurately predict churn and enable targeted retention strategies.

2. **Enhance Customer Service:**
   Proactively address issues highlighted by frequent customer service interactions to reduce churn.

3. **Tailor Retention Strategies:**
   Develop plans targeting high-usage customers and offer incentives for international plan users.

Conclusion

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● **HTML** 66.2% ● **Jupyter Notebook** 33.8%

Suggested workflows

Based on your tech stack

SLSA Generic generator Configure

Generate SLSA3 provenance for your existing release workflows

Jekyll using Docker image Configure

Package a Jekyll site using the jekyll/builder Docker image.

More workflows Dismiss suggestions