augustine-magani / **Phase-3-Project**

Code Issues Pull requests Actions Projects Wiki Security Insights Set

0 stars 0 forks 0 watching Branches Activity
Tags

Public repository

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

augustine-magani Enhance README.md with italicized emphasis	8ccea9a · 5 days ago	
Augustine_phase_3_project.i...	Add files via upload	5 days ago
Phase 3 Project - Choosing ...	Add files via upload	5 days ago
Phase 3 Project Description...	Add files via upload	5 days ago
README.md	Enhance README.md with italicize...	5 days ago
churn.csv	Add files via upload	5 days ago
utility.py	Add files via upload	5 days ago

README

SyriaTel Customer Churn Prediction

Overview

SyriaTel is a telecommunication company whose revenue model depends on recurring subscriptions. In the competitive industry, customer churn is a major threat as it reduces revenue and raises acquisition costs. Minimizing churn is therefore important for sustaining profitability and customer loyalty.

1. Business Understanding

Business Problem: SyriaTel is experiencing a high rate of customer churn, leading to significant revenue loss and reduced competitiveness. Without clear insight into which customers are most at risk and the factors driving their decisions, the company struggles to implement effective retention strategies.

Business Stakeholder: The telecom company's customer retention team is the primary business stakeholder. They directly benefit from churn predictions by acting quickly to prevent customers from leaving.

2. Data Understanding

The SyriaTel Customer Churn dataset is used for the project. The dataset includes essential customer churn attributes such as:

- State and Area code.
- International and Voice Mail Plans.
- Call rates
- Customer Service calls

The aim is to understand the structure and contents of the dataset. This involves reviewing the available features, checking their data types and identifying potential issues such as missing values or unusual patterns.

3. Data Cleaning

In this step, basic data cleaning is done to ensure the dataset is consistent and ready for modeling. The process involves:

- Checking for and handling null values
- Identifying and removing duplicate rows
- Standardizing column names by capitalizing words and separating them with underscores

4. Modelling

This section focuses on building predictive models to classify customer churn using the features in the dataset. The objective is to identify customers at risk of churn and generate insights that support effective retention strategies.

Five models are trained and evaluated:

- Logistic Regression. This is the baseline model for comparison
- Decision Tree
- Random Forest
- K-Nearest Neighbors (KNN)
- Gradient Boosting Classifier

Model performance is assessed using Recall which emphasizes the correct identification of churners. In addition, ROC-AUC which measures overall classification ability.

5. Model Evaluation

This section evaluates the performance of all trained models to determine which are most effective for predicting churn.

Key metrics such as recall and ROC-AUC are emphasized. This is because they provide the needed insight for imbalanced data. The best two models will then be selected for hyperparameter tuning.

- Gradient Boosting is the top performing model with a recall of 0.807 and an AUC of 0.912. This shows strong ability to correctly detect churners while maintaining high overall performance.
- Random Forest followed closely with an AUC of 0.908 also demonstrating high predictive power.

Gradient Boosting and Random Forest show better diagnostic ability and are the best candidates for further tuning and optimization.

Model Comparison

The tuned models Gradient Boosting and Random Forest were compared to evaluate performance on customer churn prediction. The results were as follows:

Gradient Boosting

- Accuracy: 95%
- Recall of churners: 0.81
- F1-score of churners: 0.82
- ROC–AUC: 0.923

Random Forest

- Accuracy: 91%
- Recall of churners: 0.58
- F1-score of churners: 0.65
- ROC–AUC: 0.908

Gradient Boosting achieved the best overall results especially in recall and F1 for churners; class 1. This aligns with the business objective of identifying customers likely to leave. Random Forest provided solid overall accuracy but underperformed on minority class recall.

Gradient Boosting is therefore selected as the final best model.

6. Conclusion

The project developed and evaluated different classification models to predict customer churn for SyriaTel.

Among the models tested, the Gradient Boosting Classifier provided the best balance of performance as it showed strong recall and a high ROC–AUC score. This means it is the most effective at correctly identifying customers likely to churn. This is important for business goals of reducing churn, improving customer retention and securing long term profitability.

Random Forest also performed well but its lower recall for churners makes it less suitable as the primary model.

In a nutshell, the results show the value of data driven modeling in supporting evidence based decision making for customer retention strategies.

7. Recommendations

1. Use Gradient Boosting as the main churn prediction model.
2. Use churn predictions to guide targeted retention strategies.
3. Model insights to be combined with feature analysis.
4. Frequently re-train and monitor the model.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

Jupyter Notebook 99.1% Python 0.9%

Suggested workflows

Based on your tech stack

Python application Configure

Create and test a Python application.

Python Package using Anaconda Configure

Create and test a Python package on multiple Python versions using Anaconda for package management.

Python package Configure

Create and test a Python package on multiple Python versions.

More workflows Dismiss suggestions