Dupli Checker Date: 18-05-2025

Plagiarism Scan Report

19% Plagiarism

8% Exact Match

11% Partial Match

81% Unique

Words	999
Characters	7279
Sentences	66
Paragraphs	82
Read Time	5 minute(s)
Speak Time	7 minute(s)

Content Checked For Plagiarism

1. Abstract

Customer churn analysis is very important particularly in the context of the telecom industry in order to ensure that telecommunication companies remain profitable and competitive. Applying studies to historical customer information, such as their age, frequency, bills, and communications with customer support specialists, machine learning algorithms can predict customers are likely to churn. Such models like logistic regression, random forests can accurately predict churn and can be used as effective models. From this information, the telecommunication companies can easily pin-point their churn rates and offer strategies of retention like altered offer, better services, or resolving the issues of dissatisfaction or complaints promptly so as to minimize the churn rates and improve their customer loyalty.

This report presents a comprehensive analysis and predictive modeling effort aimed at understanding and forecasting customer churn for a digital service provider. Using a dataset of over 11,000 customer records, the study applies exploratory data analysis (EDA), data preprocessing, and machine learning techniques to identify key churn indicators. The goal is to assist the business in proactively managing customer retention strategies and improving service offerings based on model insights.

2. Introduction

Introduction and Background

Customer churn is the process where users discontinue their service—is a critical metric for subscription-based businesses. High churn rates not only affect revenue but also increase the cost of acquiring new customers. Therefore, predicting customer churn allows businesses to engage at-risk users early with tailored interventions.

Problem Statement

The organization faces a challenge in identifying which customers are likely to stop using its services. Despite having access to extensive customer data, the business lacks a predictive system that can proactively flag high-risk customers. This Project aims to bridge that gap by developing a churn prediction model.

Objective of Study

- To understand the behavioural patterns and service usage that influence customer churn.
- To develop a machine learning model capable of accurately predicting churn.
- To provide actionable insights for improving customer retention.

3. Literature Review

Company and Industry Overview

The business operates in a highly competitive digital services landscape, offering subscription-based access to a variety of features. With growing customer expectations and numerous alternatives in the market, retaining customers has become more challenging.

Overview of Theoretical Concepts

Customer churn prediction typically involves supervised learning, where historical data with labelled churn outcomes is used to train classification algorithms. Techniques such as logistic regression, decision trees, and ensemble models are widely used due to their interpretability and predictive performance.

Survey on Existing Models

Several studies and industry applications leverage machine learning models to predict churn:

•Logistic Regression for interpretability.

•Random Forests and Gradient Boosted Trees for higher accuracy.

•Neural Networks in advanced settings with high-dimensional data.

Most of these models use behavioural, transactional, and demographic variables to classify churn.

4. Methodology

Scope of the Study

The study focuses on:

• Identifying opportunities to enhance customer retention.

• Using predictive modelling to forecast churn behaviour.

Data Collection Method

The dataset of a project consists of 11,260 records with 19 attributes, covering customer demographics, service usage, and financial transactions.

Visual Inspection

• Size of Dataset: 11,260 records with 19 variables.

• Data Types: Some numerical attributes are stored as object types and need conversion.

Understanding of Attributes

• Target Variable: The Target Variable in the dataset is "Churn" (Binary: 1 - Churned, 0 - Retained)

• Key Features: The key features of dataset are "Tenure," "Service_Score," "Account_Segment," "CC_Agent_Score," among others.

• Data Preprocessing: Adjustments include renaming variables, converting data types, and handling missing values.

Data Analysis Tools:

Data analysation using libraries

Data Manipulation & Computation

•pandas as pd:

Used for data manipulation and analysis. It provides data structures like Data Frames which are essential for handling structured data.

•NumPy as np:

Supports large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on them.

Data Visualization

•matplotlib.pylab as plt & matplotlib

Used for creating static, interactive, and animated visualizations in Python. pylab combines matplotlib.pyplot with NumPy for MATLAB-style plotting.

•seaborn as sns

A higher-level interface based on matplotlib that provides beautiful and informative statistical graphics.

•%matplotlib inline

A notebook magic command to display plots

Display Configuration

•pd.set_option ('display.max_columns', 100)

Ensures that up to 100 columns are displayed when printing a Data Frame, useful for wide datasets.

Statistics

•from statistics import mode

Imports the mode function to find the most frequent value in a list or array.

Machine Learning / Data Preprocessing

•from sklearn.preprocessing import *

Imports all preprocessing tools from Scikit-learn, like:

oStandardScaler: Standardizes features by removing the mean and scaling to unit variance.

oLabel Encoder: Converts categorical labels into numeric form.

•from sklearn.model_selection import train_test_split as tts

Splits datasets into training and testing sets for model validation.

Scikit-learn (sklearn)

Scikit-learn is a machine learning library that provides simple and efficient tools for data mining and analysis.

Classifiers (used to build predictive models):

•RandomForestClassifier – An ensemble method using multiple decision trees; great for handling overfitting and improving accuracy.

•GradientBoostingClassifier – Another ensemble method that builds trees sequentially to correct errors of previous ones.

•Logistic Regression – A linear model used for binary or multi-class classification.

•SVC (Support Vector Classifier) – Uses support vector machines for classification, good for high-dimensional spaces.

•KNeighborsClassifier – A simple, instance-based method; classifies data based on the majority vote of its neighbours.

•DecisionTreeClassifier – A tree-based model that splits data based on feature values.

•XGBClassifier (from xgboost) – A powerful and efficient implementation of gradient boosting; often used in winning Kaggle solutions.

Evaluation Metrics:

•classification_report – Generates a text report showing precision, recall, f1-score, and support.

•accuracy_score – Measures the ratio of correct predictions to the total number of cases.

TensorFlow and Keras

TensorFlow is a deep learning framework; Keras is its high-level API for building neural networks.Deep Learning:

•tensorflow as tf – The core TensorFlow library for numerical computation and machine learning at scale.

•keras – High-level API inside TensorFlow used to easily build and train deep learning models.

•layers (from tensorflow.keras) – Used to define different layers in a neural network

Matched Source

Similarity 2%

Title:The Next Frontier: MVNOs in the Age of 5G and AI - 6d technologies

Jun 11, 2024 · High churn rates not only affect revenue but also increase the cost of acquiring new customers. To mitigate churn, MVNOs need to enhance their customer experience through personalized services, reliable network performance, and responsive customer support.

https://www.6dtechnologies.com/blog/mvnos-in-the-age-of-5g-and-ai

Similarity 2%

Title:What is a Dataset: Types, Features, and Examples | GeeksforGeeks

Jun 6, 2024 � In a dataset, the rows represent the number of data points and the columns represent the features of the Dataset. They are mostly used in fields�...Missing: Tenure, Service_Score, Account_Segment, CC_Agent_Score,

https://www.geeksforgeeks.org/what-is-dataset

Similarity 2%

Title:A project based on Python | PDF | Mean Squared Error - Scribd

It provides data structures like Data Frames, which are essential for handling structured data. • Usage in Project: Used for data cleaning, preparation, and manipulation. Pandas is crucial for handling the time series data efficiently.

https://www.scribd.com/document/848231160/A-project-based-on-Python

Similarity 2%

Title:Developing Trading Algorithms with Python - PyQuant News

Pandas and NumPy are essential when developing trading algorithms in Python. Pandas, built on top of NumPy, provides data structures like DataFrames to handle large datasets seamlessly. NumPy, on the other hand, supports large, multi-dimensional arrays and matrices, along with a collection of mathematical...

https://www.pyquantnews.com/free-python-resources/developing-trading-algorithms-with-python

Similarity 2%

Title:

Matplotlib is a popular data visualization library in Python. It's often used for creating static, interactive, and animated visualizations in Python. Matplotlib allows you to generate plots, histograms, bar charts, scatter plots, etc., with just a few lines of code.

https://images.datacamp.com/image/upload/v1678457980/image2_6dd464d5d4.png?sa=X

Similarity 2%

Title:machine learning - Difference between standardscaler and Normalizer in ...

Aug 24, 2016 · StandardScaler standardizes features by removing the mean and scaling to unit variance, Normalizer rescales each sample.

https://stackoverflow.com/questions/39120942/difference-between-standardscaler-and-normalizer-in-sklearn-preprocessing

Similarity 2%

Title:15 Essential Python Libraries for Data Science in 2024

Aug 12, 2024 · What it is: Scikit-learn is a machine-learning library that provides simple and efficient tools for data mining and analysis.

https://medium.com/pythonforall/15-essential-python-libraries-for-data-science-in-2024-f1abe861ebdb

Similarity 2%

Title:Understanding Random Forest: The Power of Ensemble Learning

Random Forest is an ensemble learning method that combines multiple decision trees to enhance prediction accuracy and mitigate overfitting. By leveraging the strengths of individual trees, it creates a robust model suitable for both classification and regression tasks.

https://www.lyzr.ai/glossaries/random-forest

Similarity 2%

Title:Shallow Learning

Logistic Regression: A linear model used for binary or multi-class classification. It models the relationship between the input features and the predicted probabilities of different classes.

https://soulpageit.com/ai-glossary/shallow-learning-explained

Similarity 2%

Title:The Power Of Support Vector Machines: Effectiveness In High ...

Support Vector Machines offer an impressive solution for analyzing high-dimensional spaces. They effectively separate data points using support vectors and find�...Missing: •SVC | Show results with:

https://www.redshiftrecruiting.com/support-vector-machines-high-dimensional-spaces