Data Science Project predicting customer churn using Telco Customer Churn dataset

View license

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

klydia22 Update project files: remove old notebooks, add churn prediction mate...

acb5dd8 · 12 minutes ago

data	Update Phase-3_project with latest...	7 hours ago
reports	Update project files: remove old n...	12 minutes ago
src	Update Phase-3_project with latest...	7 hours ago
.gitignore	Initial commit	2 days ago
Customer_Churn_Prediction...	Update project files: remove old n...	12 minutes ago
LICENSE.md	Update Phase-3_project with latest...	7 hours ago
README.md	Update Phase-3_project, README, ...	36 minutes ago
Telco_Customer_Churn.csv	Update Phase-3_project with latest...	7 hours ago
Telco_customer_churn.xlsx	Update Phase-3_project with latest...	7 hours ago
archive.zip	Update project files: remove old n...	12 minutes ago

README License

Phase_3_Project-Telco_Customer_Churn_Prediction

Data Science Project predicting customer churn using Telco Customer Churn dataset

Project Overview

Customer Churn is one of the most significant and persistent challenges faced by telecommunication Companies, directly impacting long-term revenue and customer lifetime value. The project develops a data-driven churn prediction solution that identifies customers at risk of leaving and flag them in time for retention.

As opposed to optimizing for accuracy, the analysis is intentionally designed around business impact. The modeling approach prioritizes identifying as many true churners as possible, ensuring that the majority of the true churners are identified so that retention team can intervene proactively.

Business Problem

Telecommunication market is highly competitive in that customers can easily switch service providers. While extensive amounts of customer data are collected, many organizations still struggle to convert the collected data into actionable insights that can be meaningful in reducing churn. The project addresses the core business problem which is the lack of reliable data driven strategies that can identify customers who are about to churn. It is certain that when churners are missed; False Negatives, the company loses future revenue, making this type of error more costly than incorrectly flagging a loyal customer.

This imbalance in business cost directly informs the modeling strategy adopted in this project.

Business Objective

The primary objective of this project is to offer useful insights that can maximize customer retention. The project will:

1. Predict customer churn using historical customer behavior and service usage data

2. Maximize the identification of true churners to support proactive retention campaigns

3. Maintain model interpretability so results can be trusted and acted upon by stakeholders

4. Translate model outputs into clear, business-oriented recommendations

The primary Evaluation metric; Recall chosen to minimize costly missed churners.

Dataset Description

The analysis uses Telco Customer Churn dataset that contains detailed records of customer demographics, subscribed services, billing information and contract details. The dataset has a record of 7,043 customers and 21 features. The features include; gender, senior citizen status, tenure, contract type, payment method, internet service, streaming services, tech support, monthly charges, total charges and the target variable 'Churn' (Yes/No).

The dataset provides a realistic representation of customer behavior in the telecom industry and is well-suited for churn prediction activities.

Data Preparation

Most of the time raw data contains inconstistencies and formatting issues that can negatively impact the performance of the model. This necessesitated the need for the data to undergo several preprocessing steps to ensure data quality and reliability. The following were undertaken:

Converted TotalCharges from a string format to numeric values to enable proper mathematical operations

Handled blank strings and missing values to prevent downstream modeling errors

Droped the customerID colum, as unique identifiers do not contribute predictive value

Transformed the target variable into a binary format suitable for classification models

The steps above ensured that the dataset was both clean and analytically sound before modeling.

Feature Engineering & Preprocessing

The dataset contained a mixture of numeric and categorical variables which necessitated the implementation of a structured preprocessing pipeline to ensure consistency and prevent data leakage.

Numeric features such as tenure and charges were imputed using the median to reduce the influence of outliers and scaled using StandardScaler to ensure equal contribution to model training.

Categorical variables were transformed using One-Hot Encoding and handled carefully to account for unseen categories during testing.

Pipeline Design

A columnTransformer and Pipeline architecture was used so that all preprocessing steps were learned exclusively from the training data and applied automatically to new, unseen data. This ensures reproducibility, scalability and protection against data leakage.

Data Modeling

The project explored two complementary modeling techniques to balance interpretability and predictive power. The techiniques applied were:

1. Logistic Regression

This model served as the baseline model due to its transparency and ease of interpretation. Model coefficients provide clear insight into how each feature influences the likelihood of churn, making this model valuable for business communication.

2. Decision Tree Classifier

This model was introduced to capture non-linear relationships and interactions between features that Logistic Regression may miss. Decision Trees also provide intuitive feature importance metrics that align well with stakeholder expectations.

Both models were tuned using GridSearchCV with Recall as the optimization objective.

Handling Imbalance

The dataset exhibits a natural class imbalance with substantially more non-churners than churners. Without corrective measures, models tend to favor the majority class and underperform in identifying churrners. This issue was addressed by applying SMOTE (Synthetic Minority Oversampling Technique). Oversampling was restricted to the training dataset to avoid information leakage and SMOTE was embedded directly within the modeling pipeline.

The strategy greatly improved the model's sensitivity to churn behavior.

Model Performance

Model performance was evaluated primarily using Recall, reflecting the business priority of catching as many churners as possible.

Model Recall

Baseline Logistic Regression ~55% Tuned Logistic Regression + SMOTE ~79% Tuned Decision Tree + SMOTE ~75%

The tuned Logistic Regression model with SMOTE delivered the strongest Recall, making it the preferred solution for churn intervention.

Key Insights

The analysis of model coefficients and feature importance revealed several consistent churn drivers as follows:

Customers on month-to-month contracts are significantly more likely to churn than those on long-term contracts

Low tenure customers exhibit higher churn risk, highlighting the importance of early engagement

Fiber optic internet service is associated with elevated churn, suggesting potential service or pricing issues

Long-term contracts serve as strong retention mechanisms

The insights above provide actionable levers for customer retention strategies.

Business Impact

By shifting the modeling focus towards Recall, this project delivers tangible business value:

Improves early identification of at-risk customers

Enables targeted and cost-effective retention campaigns

Reduces long-term revenue loss associated with customer churn

Bridges the gap between data science outputs and business decision-making

Recommendations

Based on the model findings, the following actions are recommended:

1. Prioritize retention offers for customers on month-to-month contracts

2. Implement onboarding and engagement programs for new and low-tenure customers

3. Conduct deep analysis into fiber optic service dissatisfaction

4. Integrate churn prediction outputs into CRM systems for real-time intervention

Limitations

Logistic regression assumes linear relationships between features and churn. SMOTE introduces synthetic data that may not fully capture real-world behavior.

The model does not yet account for customer lifetime value or intervention costs.

Future Work

There are several opportunities existing to further enhance the solution:

i). Experiment with ensemble models such as Gradient Boosting and XGBoost

ii). Introduce cost-sensitive learning using customer lifetime value

iii). Develop time-based churn prediction models to anticipate churn windows

How to Run

Open the Jupyter Notebook included in this repository

Install required dependencies: pandas, numpy, scikit-learn, imbalanced-learn

Run all notebook cells sequentially from top to bottom

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%