Telco Customer Churn Prediction Project

Author: Abdullah Sharaf   Date: 20/05/2020

Introduction

This project aims to predict customer churn for a telecommunications company using machine learning techniques. The dataset contains information about customer demographics, services subscribed to, and billing data.

Dataset

Dataset: WA_Fn-UseC_-Telco-Customer-Churn.csv

- Rows: 7043

- Features: 21

- Target Variable: Churn (Yes/No)

Data Preprocessing

- Dropped 'customerID'

- Handled missing TotalCharges

- Converted TotalCharges to numeric

- Replaced 'No internet service' and 'No phone service' with 'No'

- Label encoding and one-hot encoding applied to categorical features

Feature Engineering

Applied multiple feature selection and transformation methods:

1. One-Hot Encoding for: InternetService, Contract, PaymentMethod

2. PCA: Reduced to 18 components explaining 96.3% of variance

3. SelectKBest: Selected top 10 features using ANOVA F-value

4. ExtraTreesClassifier: Used for feature importance ranking

5. Correlation Matrix: Removed highly correlated features

Model Training & Evaluation

Models were trained on different feature subsets and evaluated on:

- Accuracy

- Precision

- Recall

- F1-Score

Model Performance Summary:

- Logistic Regression: Accuracy = 0.79, F1 = 0.60

- Random Forest: Accuracy = 0.79, F1 = 0.56

- Gaussian Process: Accuracy = 0.80, F1 = 0.59

- SVM (RBF): Accuracy = 0.79, F1 = 0.57

- Naive Bayes (Gaussian): Accuracy = 0.71, F1 = 0.63

- Naive Bayes (Bernoulli): Accuracy = 0.77, F1 = 0.65

Conclusion

- Bernoulli Naive Bayes provided the best balance between recall and F1-score

- PCA proved effective in dimensionality reduction without significant performance drop

- Feature selection methods can influence model performance drastically

This study shows that machine learning can be effectively used to predict customer churn and potentially

Telco Customer Churn Prediction Project

Author: Abdullah Sharaf   *Date: 20/05/2020*

reduce business losses.