Churn prediction for Amazon Prime Video

Capstone Project for Machine Learning Engineer Nanodegree

Dmitry Fadeev
Munich, Germany
https://github.com/fadeetch
https://www.linkedin.com/in/dmitryfadeeff/

April 10, 2020

Abstract

Any business is naturally interested in understanding how its customers behave. Amazon Prime Video is no exception. Once initial findings are uncovered, the next step is to predict who of the customers might churn in the near future. This project fulfills two goals: a) analyses datasets from customers in Germany and the UK in order to extract valuable insights into customer behaviour and b) builds a binary classification model that predicts which customers are inactive (churned) within 6 months.

Keywords: Binary classification · Churn prediction · Ensemble

1. Problem Statement

The goal of the project is to use features in the dataset and build a binary classification model which will help identify which customers are active and inactive.

2. Datasets and Inputs

Datasets were obtained from the Amazon Data Warehouse (DWH) and contain anonymized summary statistics on around 500k customers in each DE and the UK. The files are the following:

- DE_subset.csv
- UK_subset.csv

Each file has 22 columns and around 500k rows. One row represents summary statistics on a customer level.Target variable is also provided and identifies two states: state 0 (inactive) and state 1 (active).

3. Solution Statement

Solution is represented by a classification model, which predicts whether a customer is in state 0 (inactive) or state 1 (active).

4. Benchmark Model

The field “Target” is contained in each of the datasets. The model does the classic train / test split. Therefore, it is possible to see what the overall accuracy of the model is. Also, simple logistic regression serves as a model with benchmark performance. More robust algorithm - random forest classifier scores against benchmark model.

5. Evaluation Metrics

Evaluation of the model is summarized in the ROC curve and area under the curve metric (AUC). Both benchmark and more robust models are evaluated based on AUC. Also, classification report and confusion matrix provides additional details for both models.

6. Project Design

a. Exploratory Data Analysis

This step builds distributions by specific values to get a feeling for data at hand. In particular, a researcher might be interested in understanding what is the average monetary spend of a customer depending on an engagement class. Engagement class is part of a customer segmentation model which is beyond the current project. However, in a nutshell, segmentation model scores customer behaviour in the *last twelve months* and depending on the number of transactions a customer made, assigns various labels: one-time, moderate or engaged. Also, the dataset contains other categorical variables such as content type or content age preference which might play a role in the subsequent classification model.

b. Preprocessing

i. Identify missing values

This step checks missing values in a dataframe and encodes missing values with zeros.

ii. Avoid multicollinearity

This step builds a correlation matrix and aims at identifying features with relatively high correlation. In order to avoid multicollinearity, some features are removed.

iii. Build pairplots for the feature set

This step explores how data is distributed for each quantitative feature.

iv. One-hot encoding

This step transforms categorical values into so-called one-hot-encodings so that the data can be fed into the classification algorithm.

c. Modeling

i. Split into train and test sets

This step applies functions to split the dataset into train and test subsets. The model uses 33% of the data for the test size.

ii. Fit algorithms for training sets

This step uses a library to train the algorithms. In case of convergence problems, explicit maximum number of iterations should fix the issue.

iii. Score for test sets

This step makes predictions on the test set and creates an object that contains predicted binary values which are then used for the evaluation step.

iv. Check evaluation metrics for the models

This step performs evaluation, prints accuracy, confusion matrix and ROC curve. In this step a researcher can also compare the benchmark model and other models and recommend the one with higher AUC as the final solution to the given problem.

v. Feature importance

This step seeks to make the model interpretable and figure out what is the subset of features that have the most significant outcome.

A potential next step might be to test other classification models on a set of features that exhibit the largest predictive power and compare their performance with given models. However, on a given dataset, in both UK and DE, the results with random forest classifier already provide AUC of 95%, therefore, the project did not consider any other models yet.

References:

F .Kruber, J. Wurst, “Unsupervised and Supervised Learning with the Random Forest Algorithm for Traffic Scenario Clustering and Classification”; URL https://arxiv.org/pdf/2004.02126.pdf

T. Alasalmi, J. Suutala, “Beer Classifier Calibration for Small Data Sets”, URL https://arxiv.org/pdf/2002.10199.pdf

E. Admasu, A. Teklay, “Student Performance Prediction with Optimum Multilabel Ensemble Model”; URL https://arxiv.org/pdf/1909.07444.pdf