Uniwersytet
Wrocławski

Online Shoppers Purchasing Intention

Machine Learning

Denys Tsebulia, 351322
Mafalda Costa, 351255
Mariana Carvalho, 351254

24/25

Table of contents

01 Problem statement

02 Data

03 Possible solutions

04 Work distribution

01

Problem statement

Problem statement

Online shoppers purchasing intention:

Predict whether an online shopper, based on a single session, is going to make a purchase or not.

02

Data

Data

- The dataset consists of information gathered in a period of one year from **12,330 user sessions**, such that each session corresponds to the activity of a unique user.

- Dataset of **17 features** and **one target**, the **Revenue**, which indicates if a person made a purchase or not.

- Of the 12,330 sessions in the dataset, **84.5%** (10,422) are **negative class samples**, so users that did not make a purchase, and the rest **15.5%** (1908) are **positive class samples**, so users that did make a purchase.

- **8** features (including target) in the dataset are **categorical**.

https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

Data

Features

“Administrative”, “Administrative Duration”, “Informational”, “Informational Duration”, “Product Related”, “Product Related Duration”

Numerical

These features represent the number of pages visited by the visitor in that session and total time spent in each of these page categories.

“Bounce Rate”

Numerical

Represents the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.

Data

Features

“Exit Rate”

Numerical

The percentage of pageviews on the website that end at that specific page.

“PageValues”

Numerical

Average value for a web page that a user visited before completing an e-commerce transaction.

Data

Features

“SpecialDay”

Numerical

The closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day), in which the sessions are more likely to be finalized with transaction.

“Month”, **“OperatingSystems”**, **“Browser”**, **“Region”**, **“TrafficType”**, **“VisitorType”**, **“Weekend”**

Categorical

These features are the characteristics of each user. The “month” being the month when the session happened, and the “weekend” a Boolean corresponding to the session occurring during the weekend.

Data

Features

“Revenue”

Categorical

The **target**.
Boolean that indicates whether the visitor in that session made a purchase or not.

03

Possible solutions

Existing solutions

- ***XGboost*** – 90% Accuracy
- ***LightGBM & Catboost*** – 90% Accuracy
- ***Random Forest*** – 95% Accuracy
- ***K-Nearest Neighbors*** – 91% Accuracy
- ***Support Vector Classification*** – 84% Accuracy
- ***Neural Networks (MLP, LSTM)*** – 87% (by dataset authors)

Our suggestions

- **Encode categorical features**
- **Resampling techniques**
- **Models:**
  - *Decision Tree*
  - *Support Vector Machine*
  - *Naïve Bayes*
- **Evaluation metrics:**
  - Focus on getting high accuracy and precision (minimizing false positives)

04

Work distribution

Work distribution

- **Data Analysis & Pre-processing**: Mariana + Mafalda
- **Feature Selection & Engineering**: Mafalda + Denys
- **Model building**: Mariana + Denys

References

- https://www.kaggle.com/datasets/imakash3011/online-shoppers-purchasing-intention-dataset/data
- https://www.kaggle.com/code/sasakitetsuya/clustering-and-predict-modeling-by-pycaret
- https://www.kaggle.com/code/abhishekvaishnav/eda-and-prediction#Random-Forest-Classifier
- https://www.kaggle.com/code/saifuddinlokhand/analysis-dataset-with-93-accuracy
- https://link.springer.com/article/10.1007/s00521-018-3523-0

Thank you for your attention!

:)