VYTAUTAS
MAGNUS
UNIVERSITY
Faculty of
Informatics

Customer Churn Prediction in the Software as a Service Industry

AUTHORS:

Eimantas Zaranka
eimantas.zaranka@vdu.lt

Bohdan Zhyhun
bohdan.zhyhun@vdu.lt

Milita Songailaitė
milita.songailaite@vdu.lt

Rūta Juozaitienė
ruta.juozaitiene@vdu.lt

Tomas Krilavičius
tomas.krilavicius@vdu.lt

MOTIVATION

- Customer loyalty is particularly important in areas where the revenue generation model is based on product subscription
- Attracting new customer requires much more investment than retaining existing ones, meaning analysis of customer churn is and important component for retaining wavering customers
- Analysis can also help to identify the problems that cause customers to leave

GOAL

- To create a customer churn classification model based on:
  - customer activity
  - completed orders
  - additional information
- To implement in the production workflow.

RESEARCH DESCRIPTION

1 Dataset

Dataset consisted of information on 9 215 users (6575 active and 2640 churned) covering the period from November 2020 to April 2023

Dataset contained information about:

- User
- Orders and Upselling
- User activity
- Service reviews and additional 3rd party additions to the software

2 Data Preparation

Categorical features were encoded with one hot encoding

Features with over 70 % of missing values were removed other were imputed using kNN inputer

Additional features such as days since last payment, activity since last payment and percent of days since last payment were engineered

Dataset was split into training and testing sets, where 75 % of data was used for training and 25 % for testing purposes

3 Methods

Feature selection was accomplished by using the Boruta, Boruta Shap, Correlation Coefficient and Decision Tree feature importance methods

Classification models used in this project:

- Decision Tree
- Random Forest
- Support Vector Machine
- Logistic Regression
- AdaBoost
- XGBoost

Differences between classification performance were compared via Wilcoxon and McNemar tests

EVALUATION

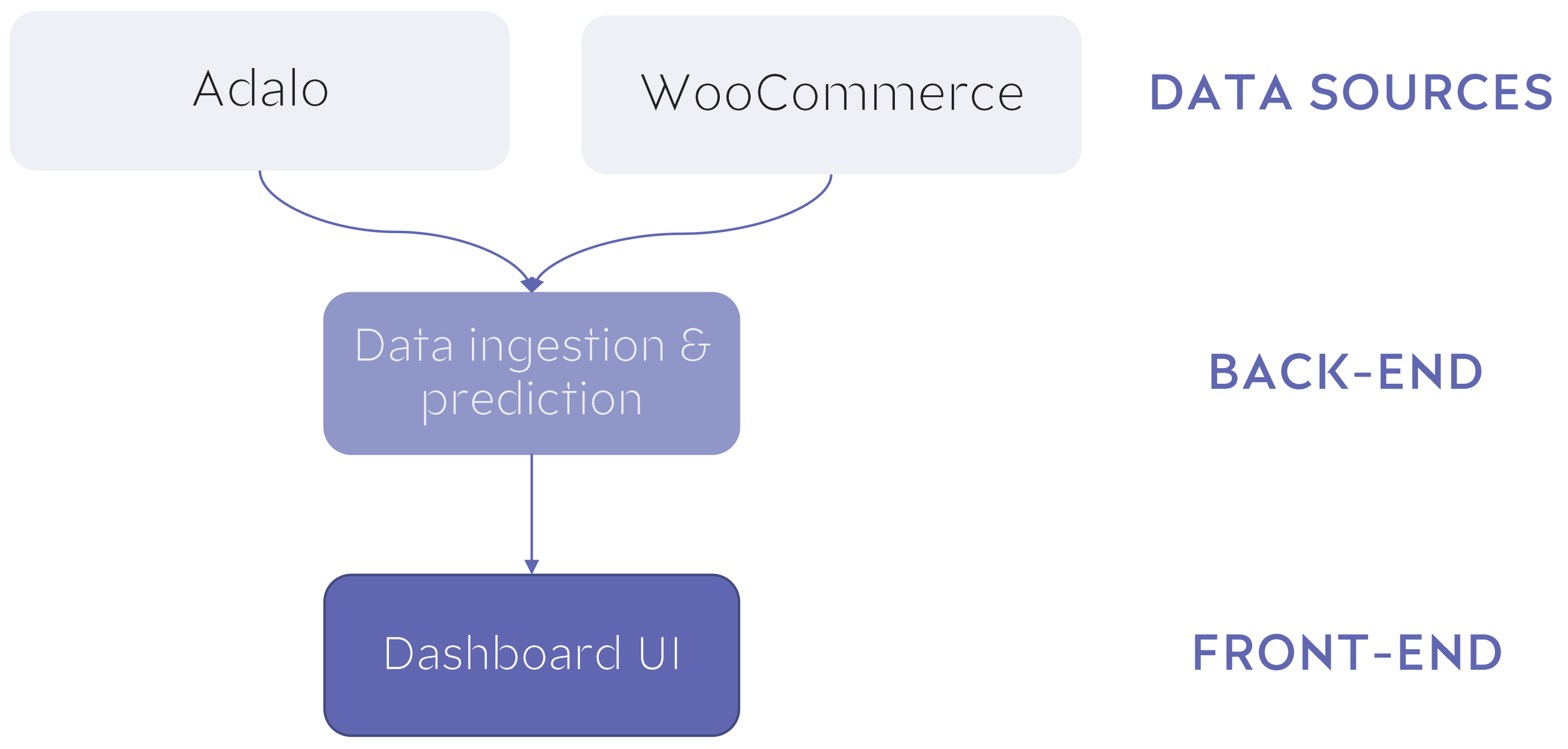
Classification models evaluation on originally provided dataset after feature selection.

Model	Accuracy	Precision	Recall	F_1
Decision Tree Classifier	0.928	0.927	0.928	0.927
Random Forest Classifier	0.935	0.935	0.935	0.935
AdaBoost Classifier	0.928	0.928	0.928	0.927
Support Vector Machine	0.910	0.909	0.910	0.909
Logistic Regression	0.890	0.883	0.890	0.889
XGBoost Classifier	**0.956**	**0.955**	**0.955**	**0.955**

Best performing XGBoost Classifier evaluation with different time horizon windows.

Time horizon window	Accuracy	Precision	Recall	F_1
7 days	0.967	0.967	0.967	0.966
14 days	0.924	0.938	0.924	0.928

PRODUCTION IMPLEMENTATION



CONCLUSIONS

- Three out of four feature selection methods identified money spent and time since last purchase as the most important features in determining customer churn.
- Best performing classification model is XGBoost Classifier with the F_1 score of 0.955
- The most promising results with one week forecast horizon achieve the F_1 score of 0.967
- The model has been tested and validated experimentally in a real-life environment to verify its effectiveness and predictive accuracy.

Co-funded by
the European Union

SustAIn
Liv
Work

CARD
CENTRE
FOR APPLIED
RESEARCH
AND
DEVELOPMENT