Churn Management
Can we predict if a customer will churn?

Timo Zaeck
Final Project
Course - "Python for Data Science", UCSD

Abstract

The E-Commerce dataset *"Ecommerce Customer Churn Analysis and Prediction"* from Kaggle was used to explore if we can use existing customer data to predict if an customer will churn. To explore this classification task the methods Logistic Regression, Decision Tree and K-Fold Cross Validation have been used. The findings show accuracy scores from 0.811 to 0.940. Depending on the model the churn rate can be predicted fairly good. But a deeper dive into the data set is necessary as the origin dataset had a lot of null values and the dependent feature "Churn" is moderate imbalanced - see slide "Limitations".

The dataset can be found here: https://www.kaggle.com/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction

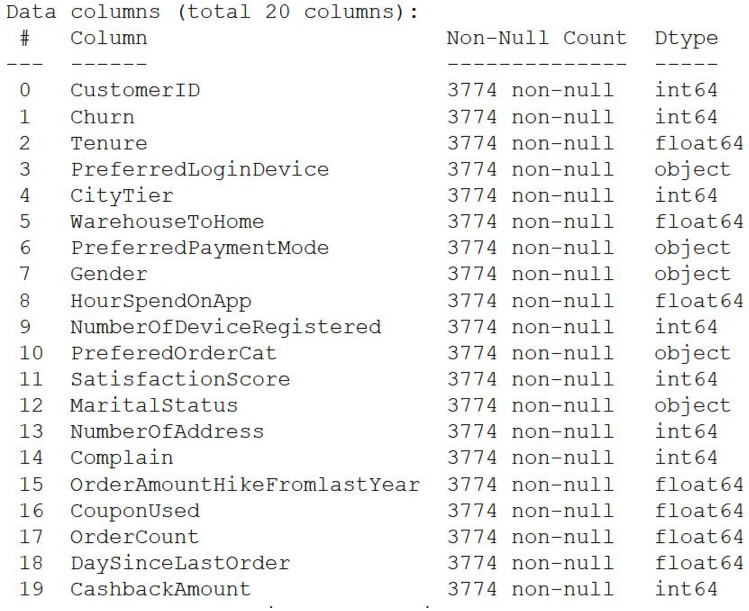
Motivation

Churn management has become increasingly important for companies as the customer acquisition costs are continuously increasing. Based on a model that could predict if an customer will churn, proactive measures could be taken to retain the customer. Insights to this could help to keep customers and reduce acquisition costs. The term “Churn” is a combination of the words “change” and “turn”. It describes the situation where a customer would like to “change” but the company would like to “turn” down his down -> “Change” + “Turn” > “Churn”. Insights to this are valuable for Sales and Marketing as well as Management but also for customers itself if the insights are used to identify features that are very important for customers and therefore could increase the service and customer satisfaction in the long term.

Dataset

The dataset used is from Kaggle. It can be found here: https://www.kaggle.com/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction

- The dataset has 5630 observations and 20 features.
- The features describe the customers and their behavior and preferences
- Features are for example “Gender”, “Marital Status”, “Order Amount hike from last year” or “Prefered Payment mode”.
- A list of all features is in figure 1 “Features”



```
Data columns (total 20 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   CustomerID                   3774 non-null   int64  
 1   Churn                        3774 non-null   int64  
 2   Tenure                       3774 non-null   float64
 3   PreferredLoginDevice         3774 non-null   object 
 4   CityTier                     3774 non-null   int64  
 5   WarehouseToHome              3774 non-null   float64
 6   PreferredPaymentMode         3774 non-null   object 
 7   Gender                       3774 non-null   object 
 8   HourSpendOnApp               3774 non-null   float64
 9   NumberOfDeviceRegistered     3774 non-null   int64  
 10  PreferedOrderCat             3774 non-null   object 
 11  SatisfactionScore            3774 non-null   int64  
 12  MaritalStatus                3774 non-null   object 
 13  NumberOfAddress              3774 non-null   int64  
 14  Complain                     3774 non-null   int64  
 15  OrderAmountHikeFromlastYear  3774 non-null   float64
 16  CouponUsed                   3774 non-null   float64
 17  OrderCount                   3774 non-null   float64
 18  DaySinceLastOrder            3774 non-null   float64
 19  CashbackAmount               3774 non-null   int64  
```

Figure 1: Features

Data Preparation and Cleaning

- **Checking for null values:**
  - Several features had null values. All entries with null values had been dropped.
  - After this step the number of observations dropped from 5630 to 3774
- **Checking for duplicates:** No duplicates have been found.
- **Checking for outliers:**
  - Using Box-Plot method. Calculate Interquartile Range (IQR). Outliers are defined as the observations that are below Q1 − 1.5 x IQR or above Q3 + 1.5 x IQR
  - Outliers removed from these features: “Tenure”, “Warehouse to home”, “Order Amount Hike from last year”, “Coupon Used”, “Order Count”, “Day since last order”, “Cashback Amount”
  - After removing outliers 2856 observations had been left.
- **Dummy variables:** Switching five categorical features to numerical features:
  - “PreferredLoginDevice”, “PreferredPaymentMode”, “Gender”, “PreferedOrderCat”, “MaritalStatus”
- **Scaling:** Features have been scaled with StandardScaler

Research Question

1. How well can we use existing customer data to predict if the customer will churn using three different classification models: Logistic Regression, Decision Tree Classifier and K-Fold Cross Validation?

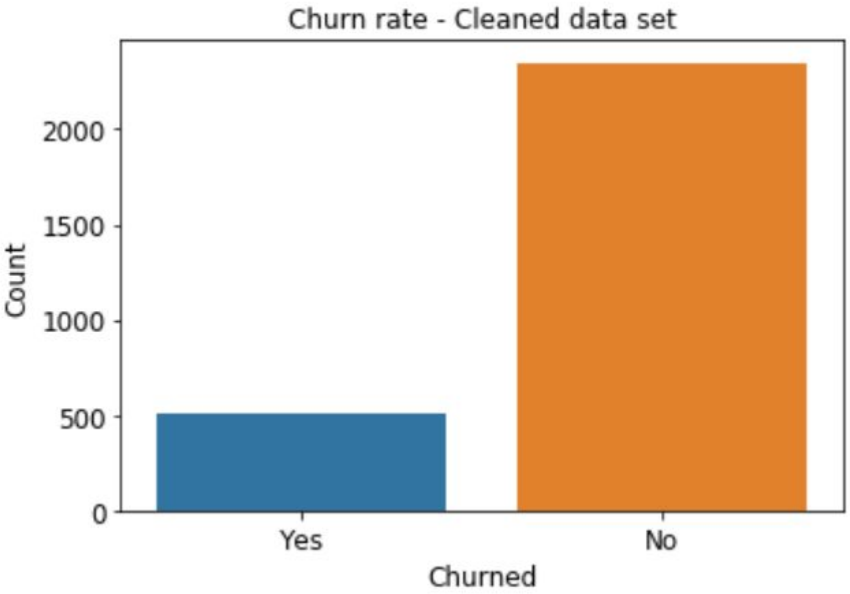
Methods

- As we have a classification task we use the methods Logistic Regression, Decision Tree Classifier and K-Fold-Cross-Validation
- Logistic Regression and Decision Tree Classifier:
    - Cleaned and prepared dataset has been splitted in a test and training data set with a ratio of 0.7 for training data set and 0.3 for test data set.
    - For evaluation accuracy score and confusion matrix have been used
- For K-Fold-Cross-Validation ten different settings have been used.
    - Folds = 20, n neighbors = 2,3,4,5,6
    - Folds = 40, n neighbors = 2,3,4,5,6
    - A sum of the accuracy scores had been build and the mean calculated to evaluate the performance of K-Fold-Cross-Validation

Findings - Explorative Data Analysis

- The churn rate in the cleaned dataset

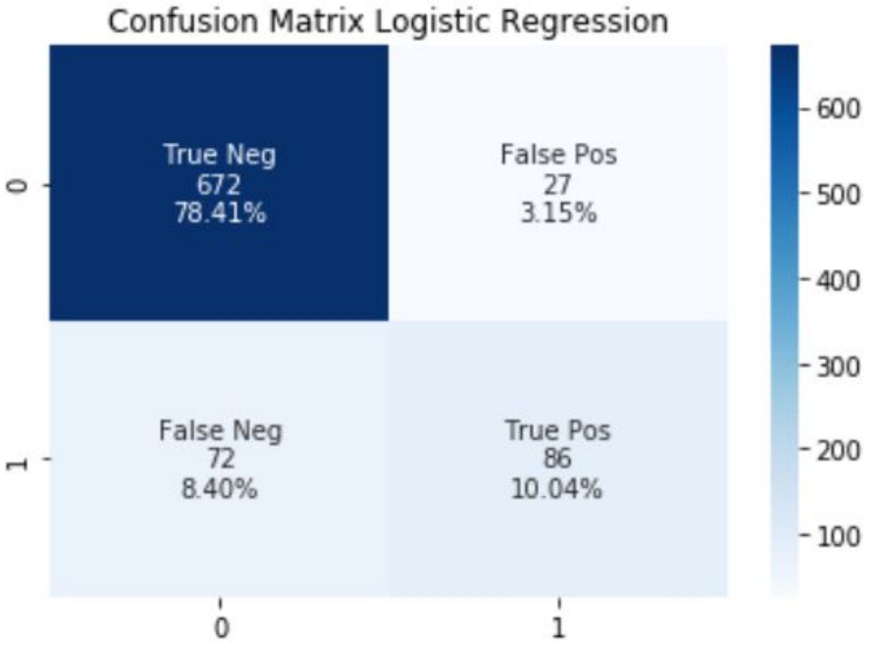
Churned	Count	%
Yes	507	17,75
No	2349	82,25


Churn rate - Cleaned data set

Findings - Logistic Regression

- Accuracy score: 0.88
- Confusion matrix:
    - 672 true negative
    - 86 true positive
    - 72 false negative
    - 27 false positive.

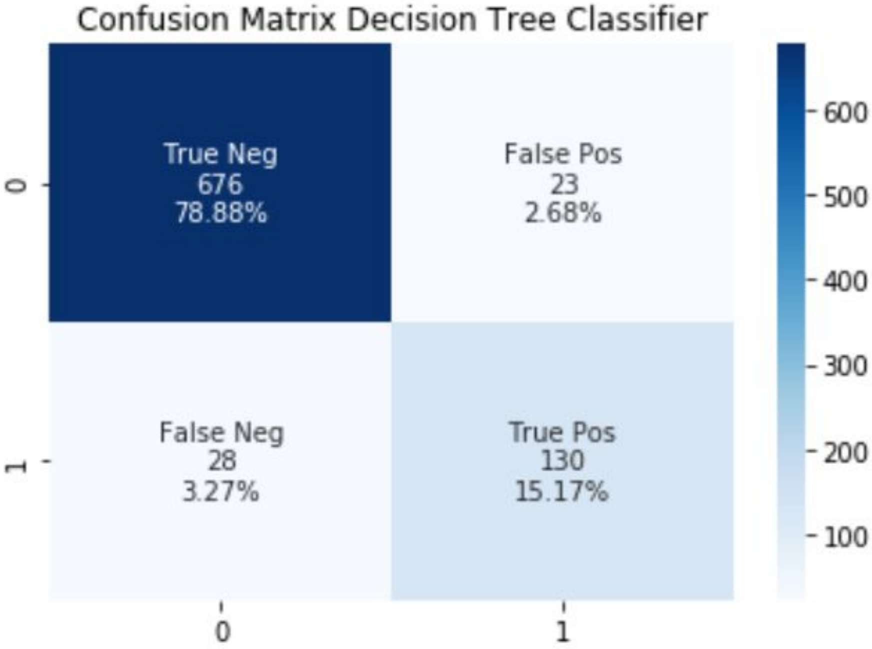
The model is able to predict a high number of true negative and true positive. But the number especially for false negative seems pretty high.



Findings - DecisionTreeClassifiern

- Accuracy score: 0.94
- Confusion matrix:
  - 676 true negative
  - 130 true positive
  - 28 false negative
  - 23 false positive

The model is able to predict a high number of true negative and true positive. The number especially for false negative is far lower compared to Logistic Regression.



Findings - K-Fold Cross validation

Setting	Folds	N neighbors	Mean of accuracy score
#1	20	2	0.875
#2	20	3	0.816
#3	20	4	0.857
#4	20	5	0.874
#5	20	6	0.855
#6	40	2	0.874
#7	40	3	0.811 = lowest accuracy
#8	40	4	0.859
#9	40	5	0.879 = highest accuracy
#10	40	6	0.857

Findings - Comparison of the methods

	Logistic Regression	**Decision Tree Classifier**	**K-Fold-Cross validation**
Accuracy	0.88	0.94	From 0.811 to 0.879
Confusion matrix	- 672 true negative - 86 true positive - 72 false negative - 27 false positive.	- 676 true negative - 130 true positive - 28 false negative - 23 false positive	

- Decision Tree Classifier has the highest accuracy and far less “false negatives” than Logistic Regression
- K-Fold-Cross validation has the lowest accuracy score

Limitations

- The model can only be used for this data available. It can not be transferred to other customer structures.
- The data set already contained different variables but missed some customer information that might be useful to improve the model for example like age, total money spent or salary.
- The way the null values and outliers have been handled here had reduced the observations available for the model from 5630 to 2856 (see slide “Data Preparation and Cleaning”). Deeper knowledge is needed to use more advanced cleaning techniques to avoid such a huge loss of observations.
- The feature “Churn” has an moderate imbalanced data with a ratio of 17.75 to 82.25. Deeper knowledge is needed to handle this properly with more advanced techniques - like Downsampling and Upweighting. See “Imbalanced Data”, Google, https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

Conclusions

- From the methods chosen to predict if a customer will churn Decision Tree Classifier has the best result. This can help Sales and Marketing to take action to prevent a customer from churning
- The error rate for false positives and false negatives in the confusion matrix especially for Logistic Regression seems very high. This could lead to costs for activities to prevent a customer from churning while the customer does not want to churn.
- To further improve the prediction new features describing the customer could be helpful. It should be discussed within Sales & Marketing.
- Further improvement of the data preparation and cleaning is needed with techniques that go beyond this course.
    - Better handling the imbalance of “Churn” is needed
    - Better handling of null values and outliers to avoid a huge drop in observations

Acknowledgements

I got the Data from Kaggle.
https://www.kaggle.com/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction

References

- The code for the visualisation of the confusion matrix has been adapted from this source: https://medium.com/@dtuk81/confusion-matrix-visualization-fc31e3f30fea
- Imbalanced Data - Google https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data