DIE ZEIT

Churn Prediction

Beat the logistic regression model of 'Die ZEIT'

Author: Sebastian Fuhrer

TABLE OF CONTENTS

01

THE COMPILATION OF THE DATA SET

THE COMPILATION OF THE DATA SET

BASIC UNIT

The subscription orders for ZEIT and ZEIT Digital from the years > 2012 from households with a max of 4 subscriptions were considered

EXPLANATORY VARIABLES

Some examples: purchase completion, payment behavior, subscription and object form, ZEIT shop, demographics, newsletter and email performance

DIMENSIONS

171 features

From 175.130 customers

Target variable 'churn'

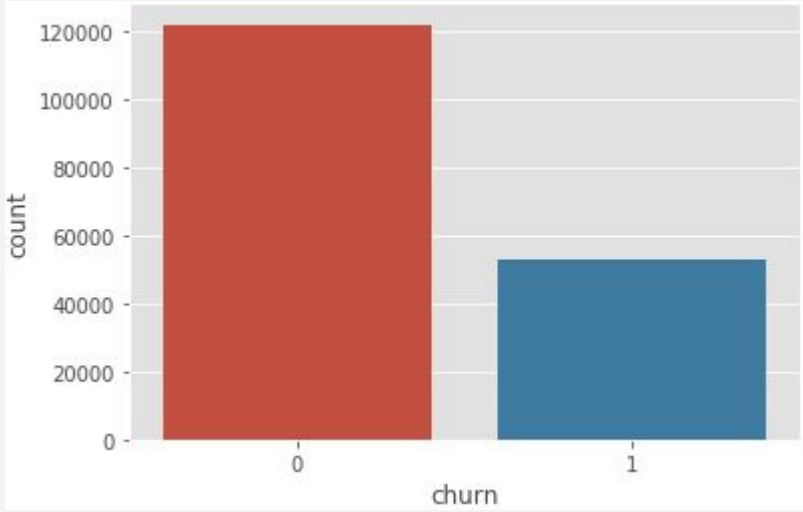
Specialty

A model already exists. A regression approach with Lasso / Ridge feature selection

02

EXPLORATORY DATA ANALYSIS

CHURN



CUSTOMER BEHAVIOR

The distribution of the target variable ‘churn’ is unbalanced. About 52.905 customers churned and 122.140 renew or still have a subscription from ‘Die ZEIT’ in the period considered.

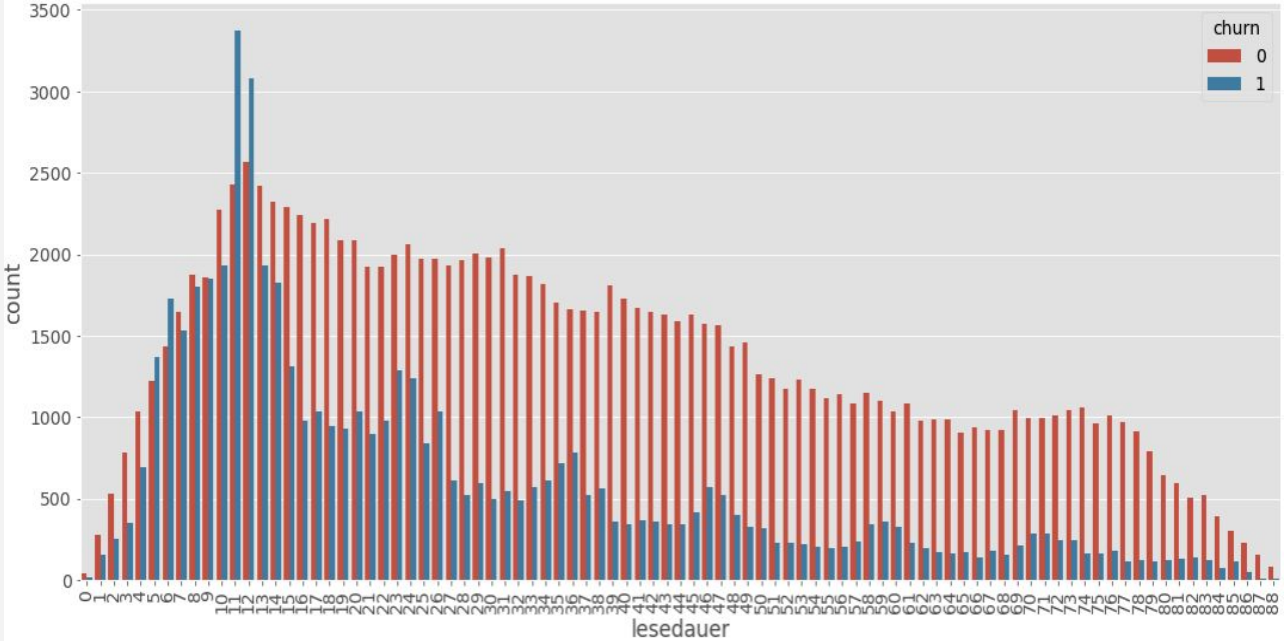
CHURN AND RETENTION RATE

69,78%
RETENTION

30,22%
CHURN

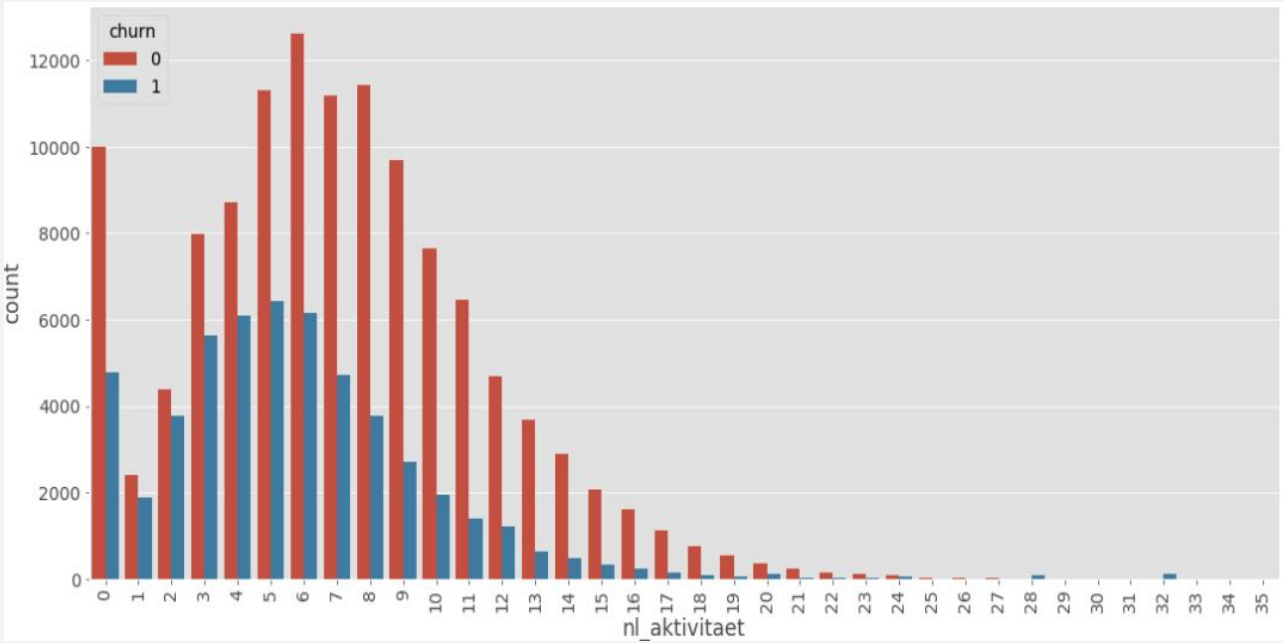
READING TIME

- The churn rate decreases with increasing reading time
- The highest churn rate is after 11 or 12 months
- 12 months of reading seems to be a threshold towards loyal customers
- There are always smaller peaks in the churn rate, after two years, three years, etc.



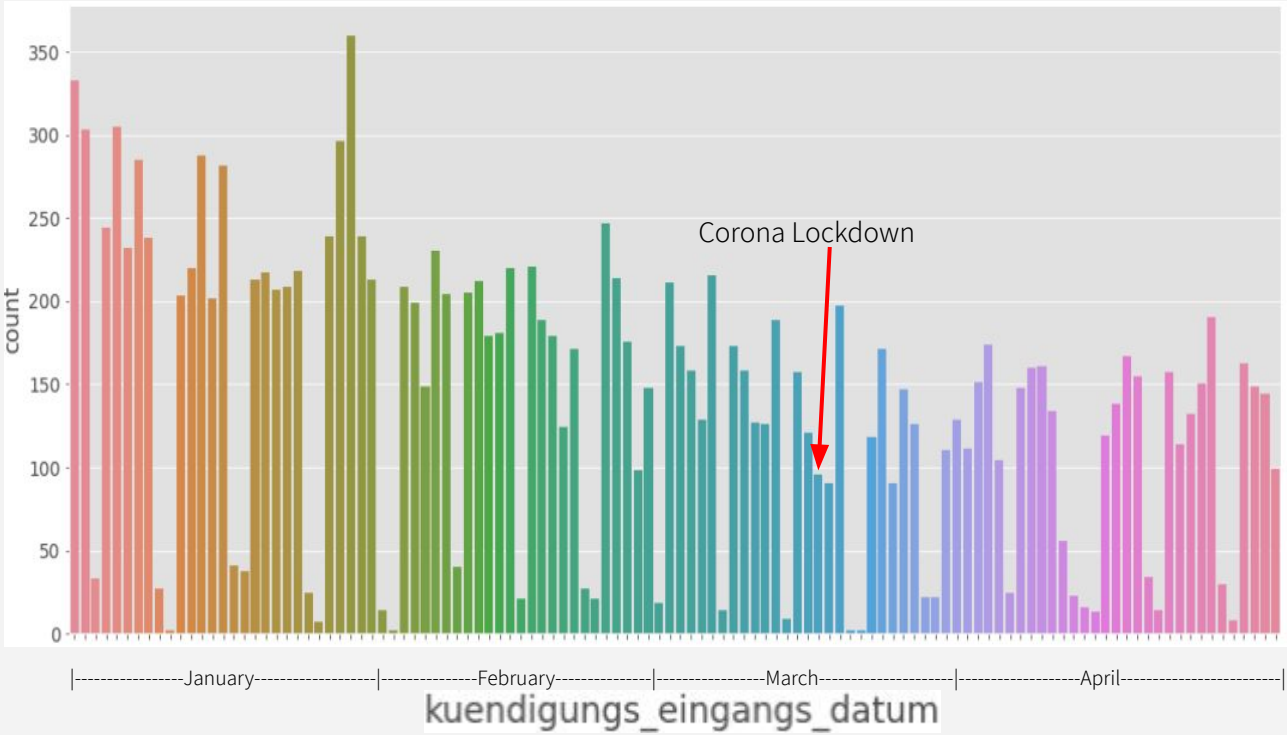
NEWSLETTER ACTIVITY

- The newsletter activity in the households is a strong predictor
- If the household in question has more than 5 active newsletters, its churn rate drops sharply.



DOES CORONA HAVE AN EFFECT ON THE CHURN RATE?

- We do not have a comparable period, as we are only looking at one year
- At the end of the year, at the beginning of the next year, there is increased churn activity
- Two days after the lockdown on March 18, 2020, there is a small peak in the churn rate

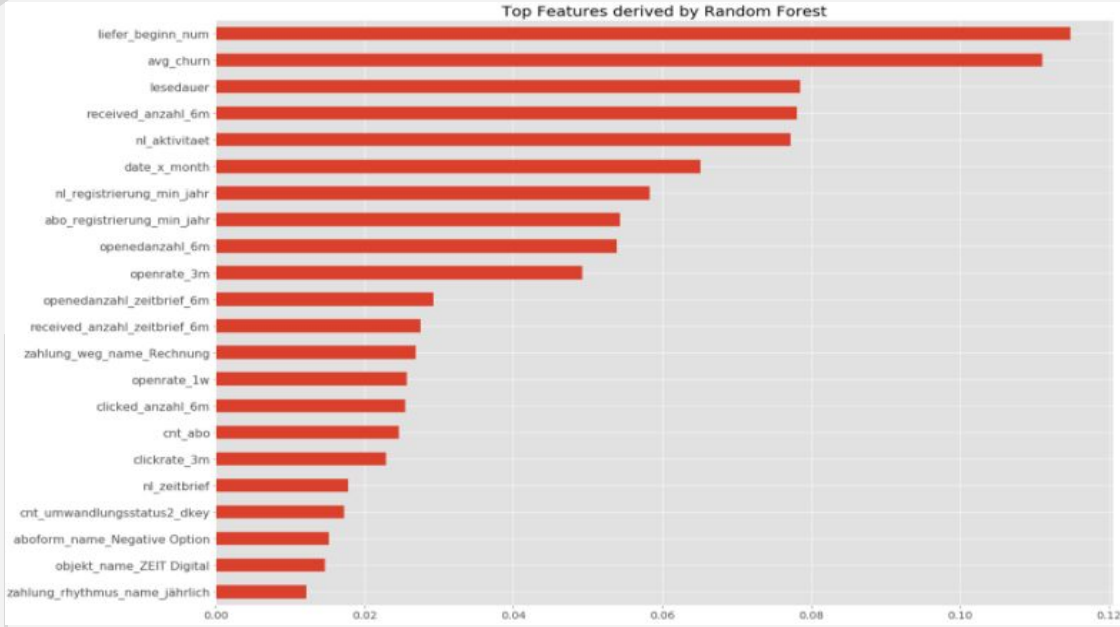


03

FEATURE IMPORTANCE

FEATURE SELECTION RANDOM FOREST

We had to drop the top feature 'avg_churn' because of data leakage. Delivery start, reading time, subscription registration and newsletter registration are top explanatory features.



TOP FEATURES

11,3%
DELIVERY START

7,9%
READING TIME

TOP 20 FEATURES

After checking the multicollinearity between the features, these top features remain

Features

- zahlung_weg_name_Rechnung: Payment method, in particular invoice payment method
- zahlung_rhythmus_name_jährlich: Payment frequency, especially annually
- aboform_name_Negative Option: Subscription form especially negative option
- objekt_name_ZEIT Digital: Object especially ZEIT Digital
- nl_zeitbrief: Newsletter Zeitbrief
- liefer_beginn_num: Delivery start
- nl_aktivitaet: Newsletter activity
- abo_registrierung_min_jahr: The year in which the first subscription was ordered
- nl_registrierung_min_jahr: First registration of the newsletter
- received_anzahl_6m: Number of newsletters received in 6 months
- date_x_month: Day of termination and random date
- openedanzahl_6m: Number of newsletters opened in 6 months
- openrate_3m: How many times was the EMail opened in 3 months
- received_anzahl_zeitbrief_6m: Number of newsletter Zeitbrief received within 6 months
- openedanzahl_zeitbrief_6m: Number of open newsletter Zeitbrief within 6 months
- clicked_anzahl_6m:: Number of clicks within 6 months
- cnt_abo: How many additional subscriptions per household
- cnt_umwandlungsstatus2_dkey: Number of additional subscriptions that also converted (not just free trial period)
- openrate_1w: Number of opened EMails in a week
- clickrate_3m: Number of clicks within 3 months

04

MACHINE LEARNING MODELS

WHAT IS OUR GOAL?

Every model has a threshold that discerns positives from negative predictions. The lower the threshold the more instances get predicted positive. We can tweak the machine learning models by higher or lower the threshold.

PRECISION

What is the percentage of correctly predicted positives? As a percentage: How often did the model predict positives?

ACCURACY

How often the model has been right? When one class is very rare it leads to false conclusions.

RECALL

How many of the positives did the model correctly predict? The recall is also called sensitivity.

F1-SCORE

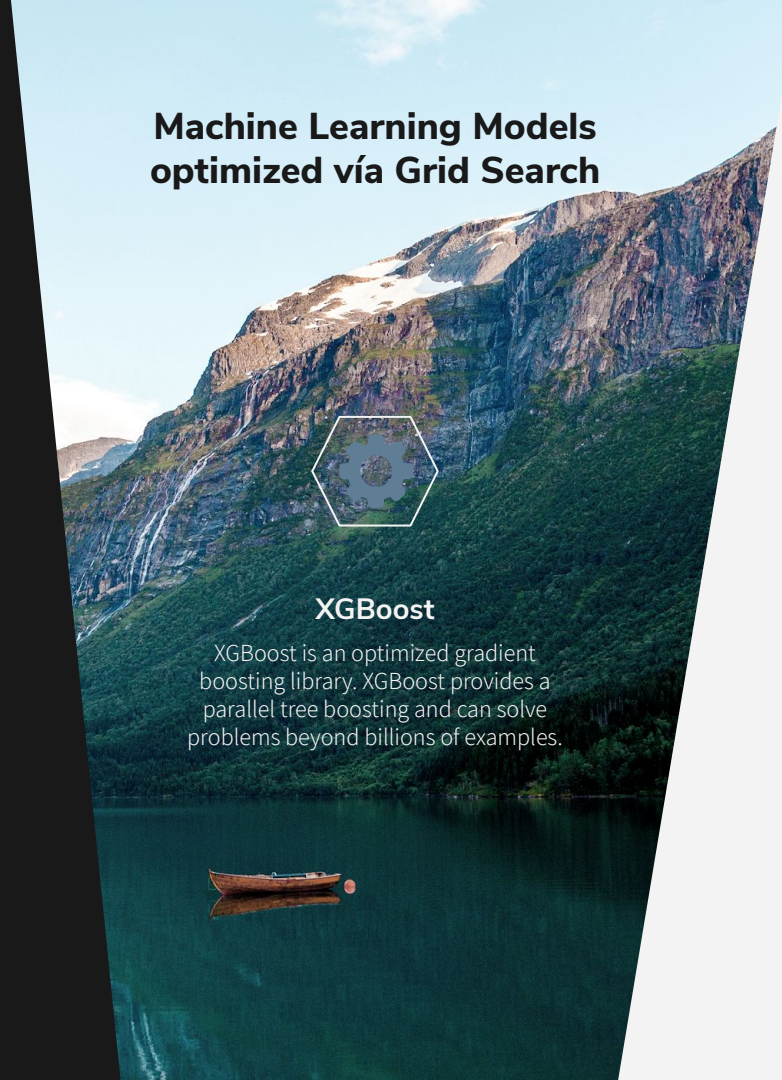
Jupiter is the biggest planet in our Solar System and the fourth-brightest object in the sky

RESULTS ON TRAINING DATA SET

- Three of the machine learning models stand out in terms of the results on the training data set
- The PRECISION values are very high with up to 93% and even the RECALL values are astonishingly high with at least 81%, which could be a signal for overfitting
- On the next page we optimize three models via grid search and look at the results on the test data set

	Model	Accuracy	Recall	Precision
0	logreg	0.66	0.67	0.66
1	tree	0.65	0.66	0.64
2	forest	0.92	0.94	0.90
3	XGB	0.68	0.66	0.69
4	AdaBoost	0.67	0.68	0.67
5	KNN	0.75	0.81	0.72
6	SVM	0.70	0.68	0.70
7	Stacking	0.92	0.89	0.93

Machine Learning Models optimized vía Grid Search



KNN

KNN classifies based on similarity measures. KNN is a non-parametric algorithm.

XGBoost

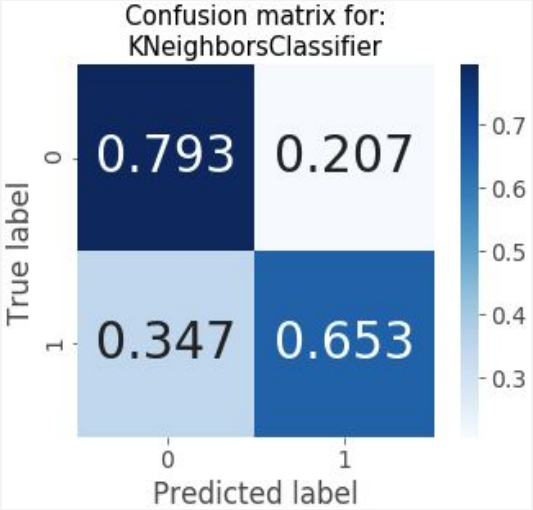
XGBoost is an optimized gradient boosting library. XGBoost provides a parallel tree boosting and can solve problems beyond billions of examples.

Random Forest

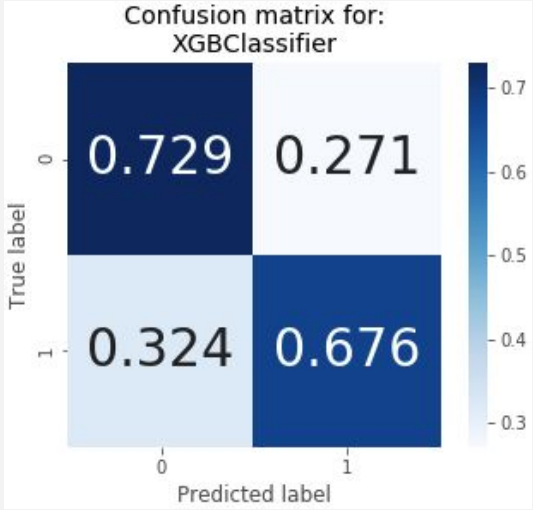
The random forest is a supervised learning algorithm that randomly creates and merges multiple decision trees into one forest.

RESULTS ON TEST DATA SET

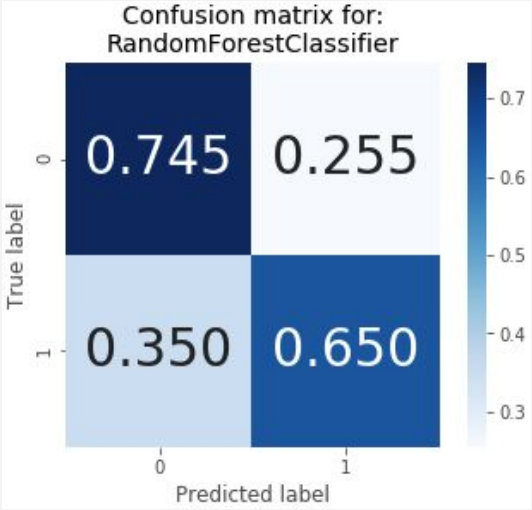
KNN



XGBoost



Random Forest

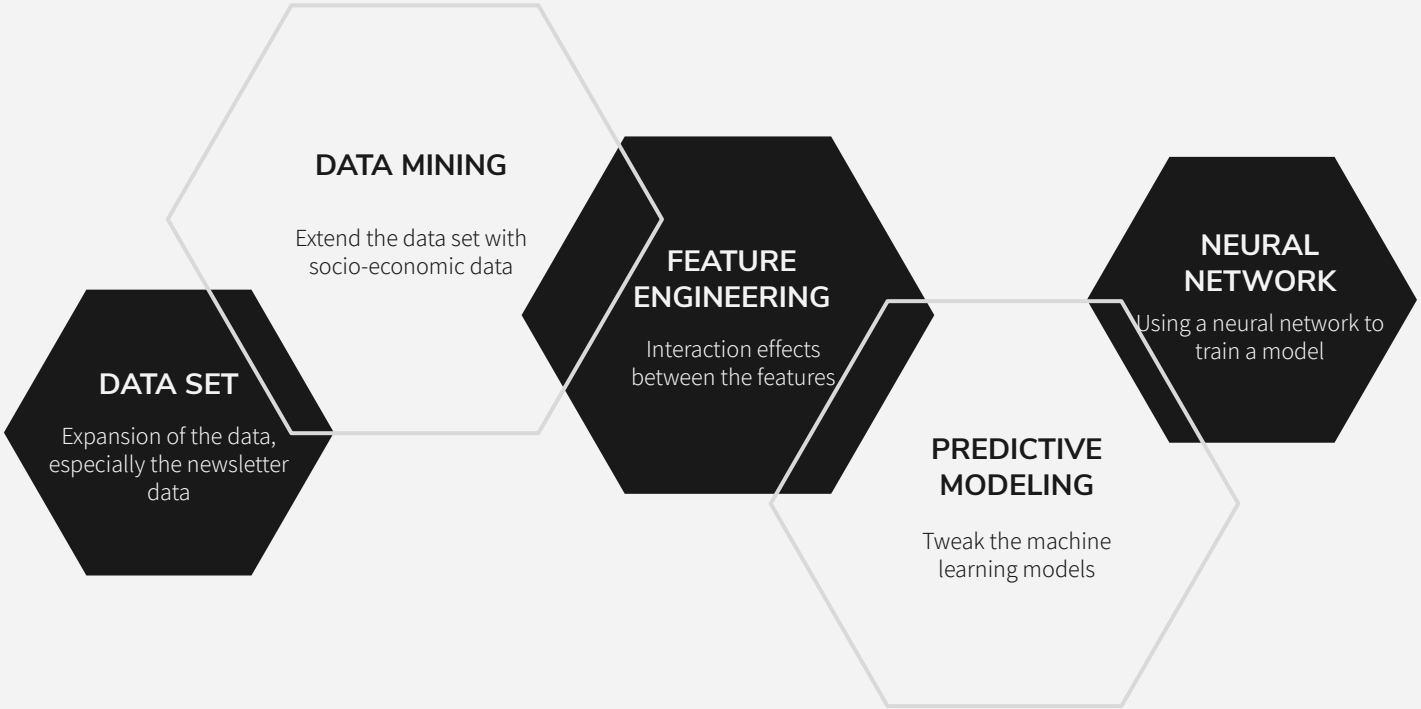


ROC AUC Score: **0,72** **0,70** **0,69**

05

FUTURE WORK

FUTURE WORK


DATA SET
Expansion of the data, especially the newsletter data
DATA MINING
Extend the data set with socio-economic data
FEATURE ENGINEERING
Interaction effects between the features
PREDICTIVE MODELING
Tweak the machine learning models
NEURAL NETWORK
Using a neural network to train a model

THANKS

Does anyone have any questions?

fuhrer_sebastian@web.de
+49 174 6363886
https://github.com/booty7