DIE ZEIT

Churn Prediction

Beat the logistic regression model of 'DIE ZEIT'

Author: Sebastian Fuhrer, Marco Zagermann

WHAT IS CHURN PREDICTION?

PROBLEM

A common problem of many newspapers and magazines: Subscribers may end their subscription (“churn”)

CONSEQUENCES

- Loss of Sales
- New customers have to be recruited for compensation (difficult)

SOLUTION

- Predict churn with AI
- Prevent churn from happening (E-Mail, special offer, etc.)

TABLE OF CONTENTS

01

THE DATA SET

THE DATA SET

	Unnamed: 0	auftrag_new_id	liefer_beginn_evt	kanal	objekt_name	aboform_name	zahlung_rhythmus_name	lesedauer	rechnungsmonat
0	3	25B535B7-D3F9-4804-9FFE-A8C813C8A593	2013-10-17	andere	ZEIT Digital	Festabo	jährlich	72	0
1	4	C348C873-8229-4F4F-8B50-5B4C06B2E872	2013-09-12	andere	ZEIT Digital	Festabo	vierteljährlich	70	0

BASIS

- Subscription orders for ZEIT and ZEIT Digital since 2012
- Max. four subscriptions per household

VARIABLES

- subscription details
- demographics
- geographical info
- temporal features
- newsletter & email activity

DIMENSIONS

- 171 features
- 175,130 customers
- Target variable 'churn'

CHALLENGE

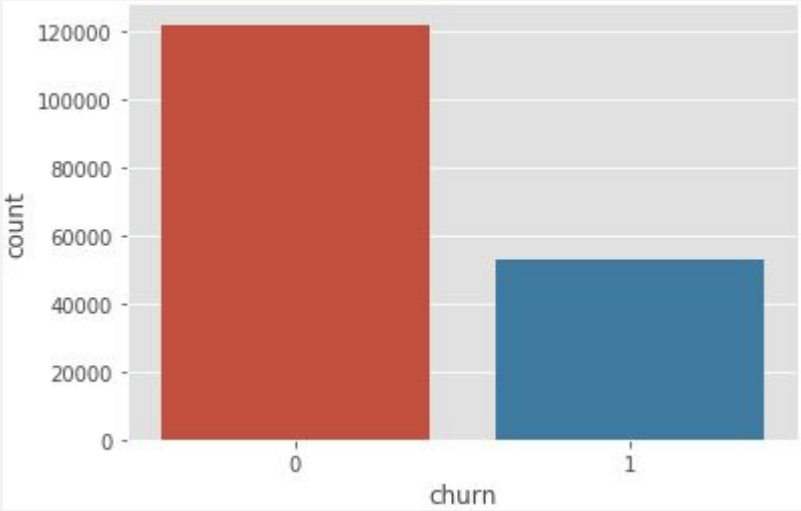
Pre-existing model:

- Logistic regression
- Lasso / Ridge feature selection

02

EXPLORATORY DATA ANALYSIS

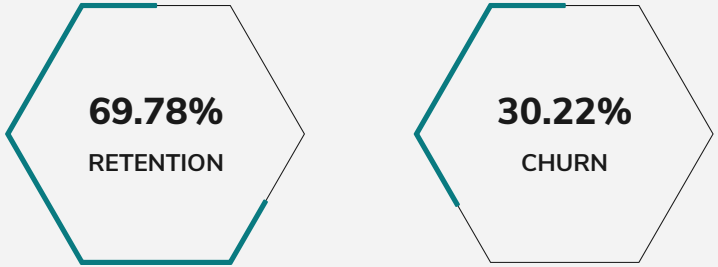
CHURN



CUSTOMER BEHAVIOR

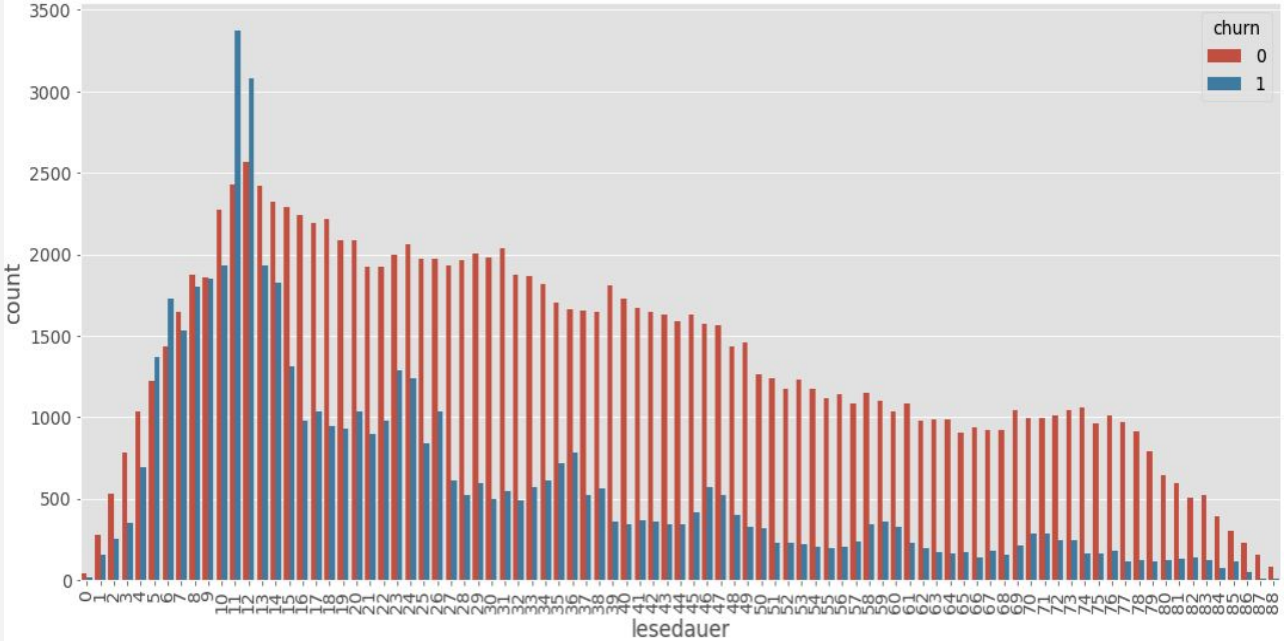
- Unbalanced distribution of target variable ‘churn’
- 52,905 customers churned
- 122,140 did not churn in the period considered (06/2019-05/2020)

CHURN AND RETENTION RATE



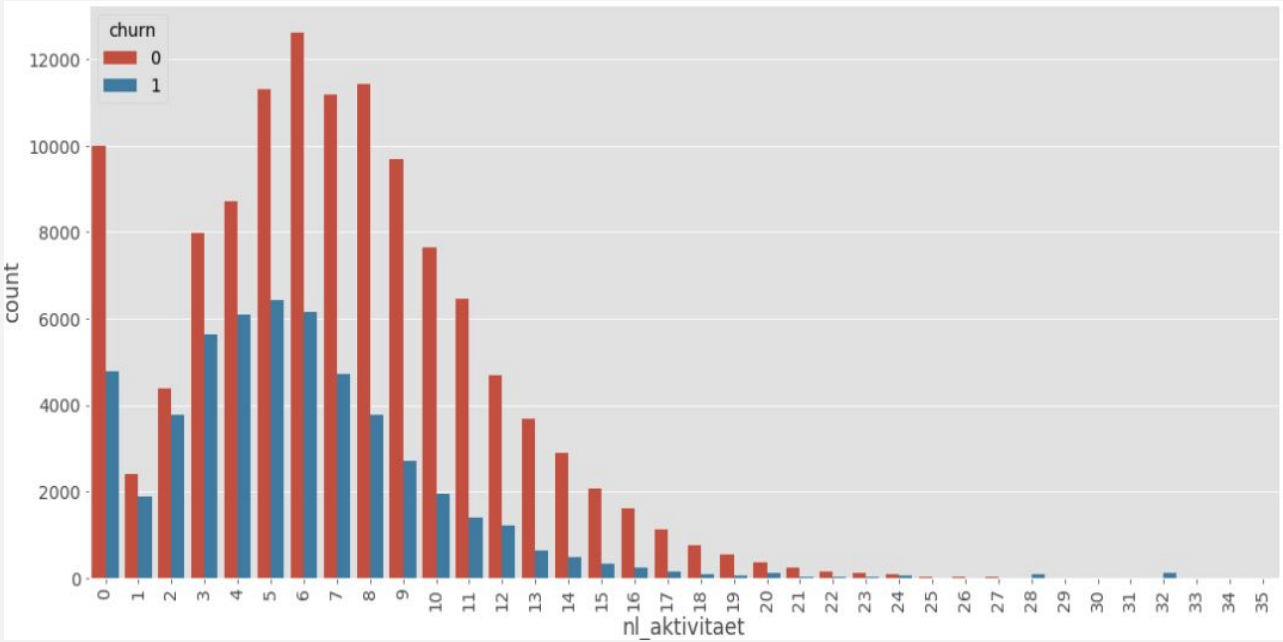
MONTHS OF READING

- Churn rate decreases with increasing reading time
- 12 months of reading is the major threshold towards loyal customers
- Smaller peaks in the churn rate, after two years, three years, etc.



NEWSLETTER ACTIVITY

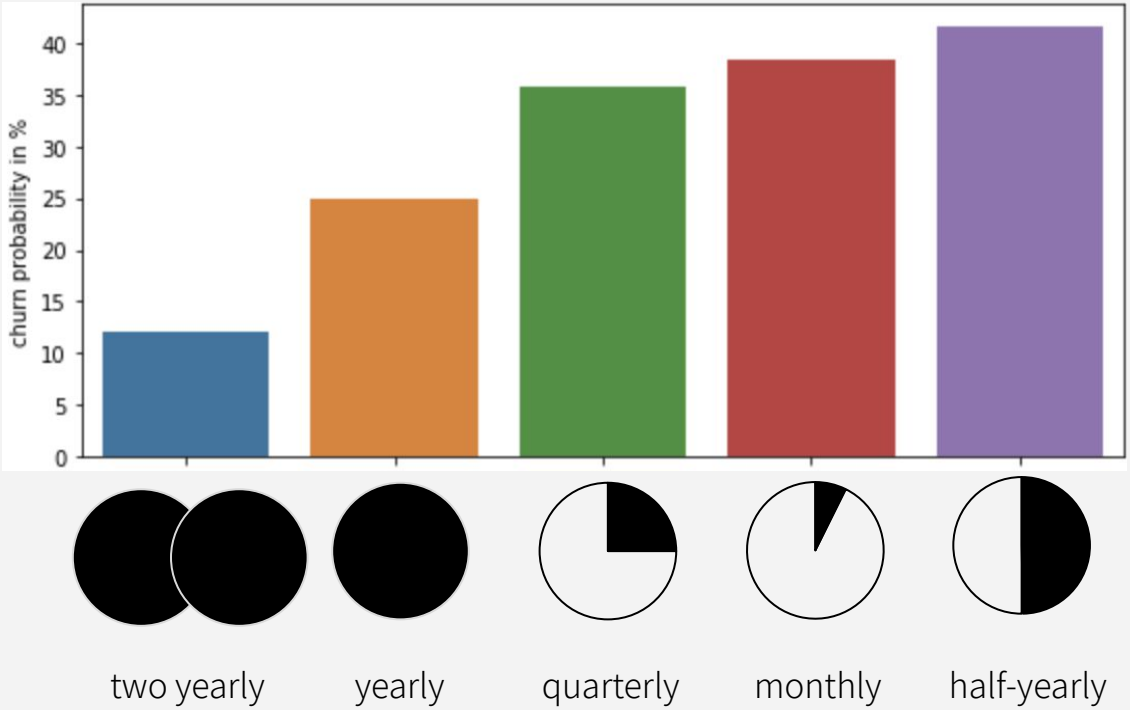
- Newsletter activity is a strong predictor
- For more than two newsletters per household the churn rate drops strongly



DIGITAL VS. PAPER VS. CHRIST & WELT

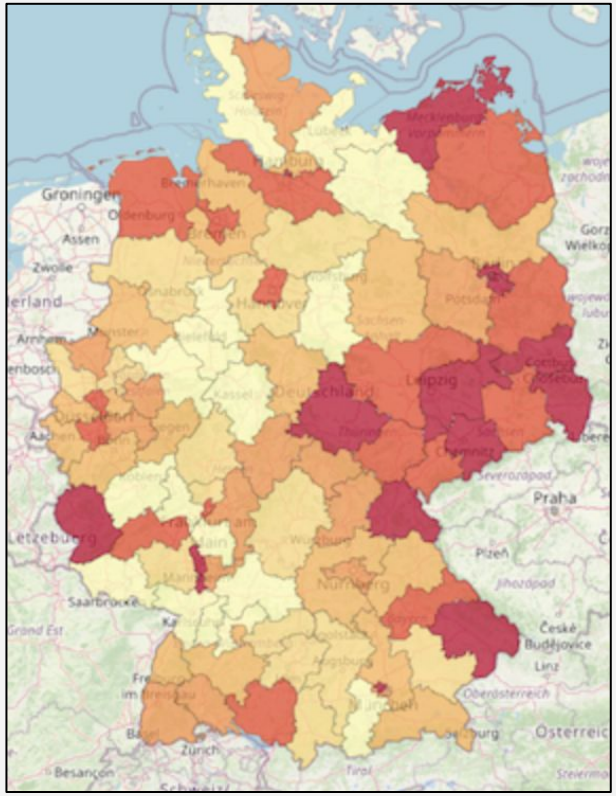
Christ & Welt | Digital | Paper

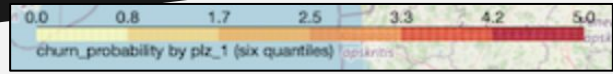
RHYTHM OF PAYMENT

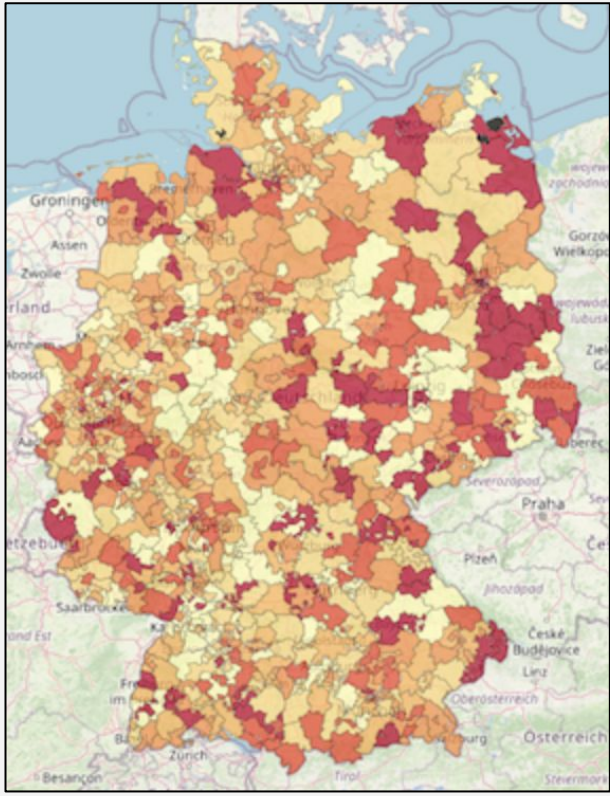


POSTAL CODE


Groningen
Assen
Zwolle
Hamburg
Bremen
Berlin
Potsdam
Hannover
Leipzig
Praha
Plzeň
Nurnberg
Saarbrücken
České Budějovice
Linz
Augsburg
München
Österreich
Zürich
Besançon

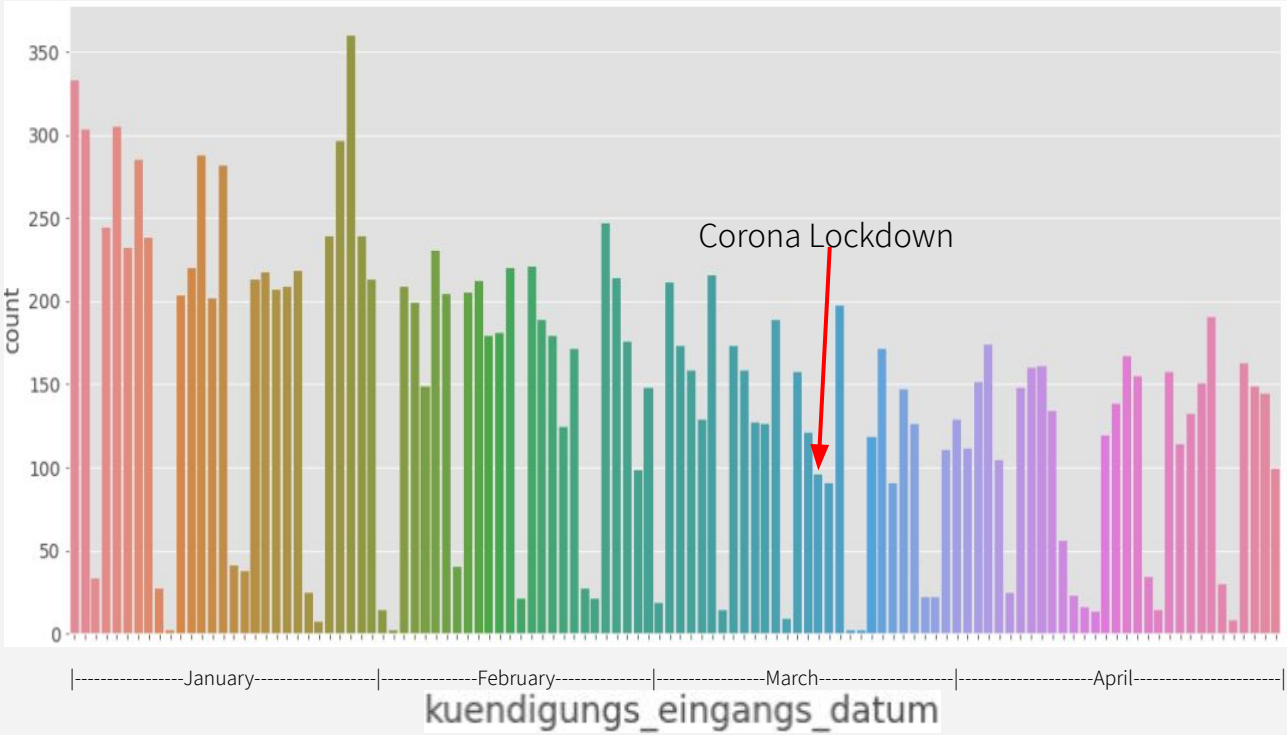

Groningen
Assen
Zwolle
Praha
Plzeň
Saarbrücken
České Budějovice
Linz
Österreich
Zürich
Besançon


0.0
0.8
1.7
2.5
3.3
4.2
5.0
churn_probability by plz_1 (six quantiles)


Groningen
Assen
Zwolle
Praha
Plzeň
Saarbrücken
České Budějovice
Linz
Österreich
Zürich
Besançon

DOES CORONA HAVE AN EFFECT ON THE CHURN RATE?

- We do not have a comparable period, as we are only looking at one year
- Increased churn activity at the end or beginning of the year
- No obvious change due to Corona Lockdown



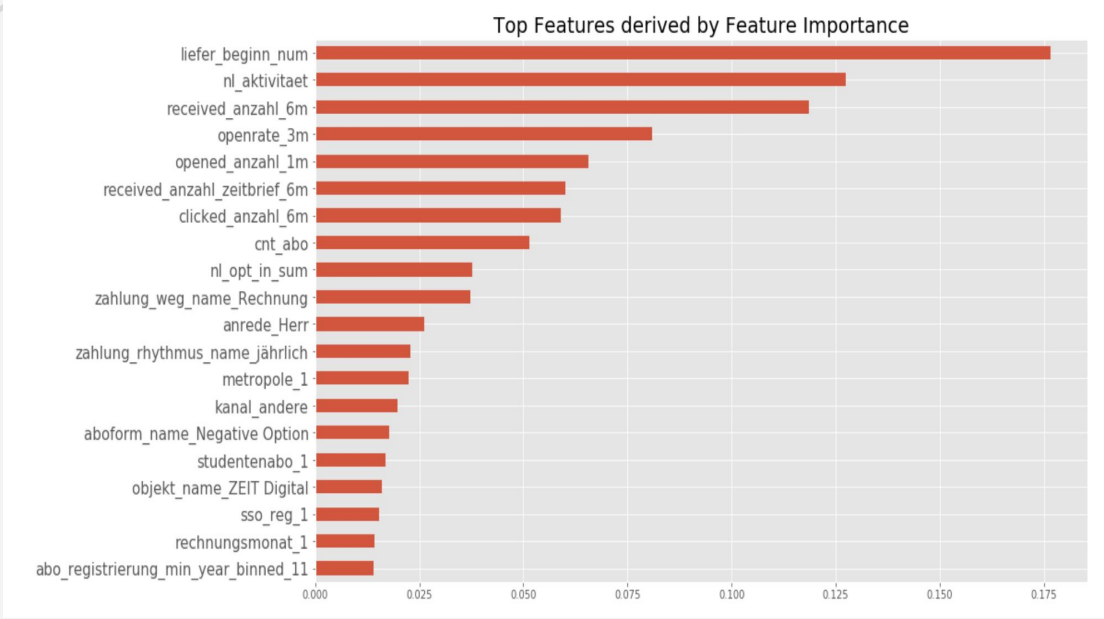
03

FEATURE SELECTION

THREE ALGORITHMS FOR FEATURE SELECTION

1. Correlation with churn
2. SelectKBest from Scikit-Learn
3. Feature importance from decision trees

→ Several different feature sets with 20 or 30 features each

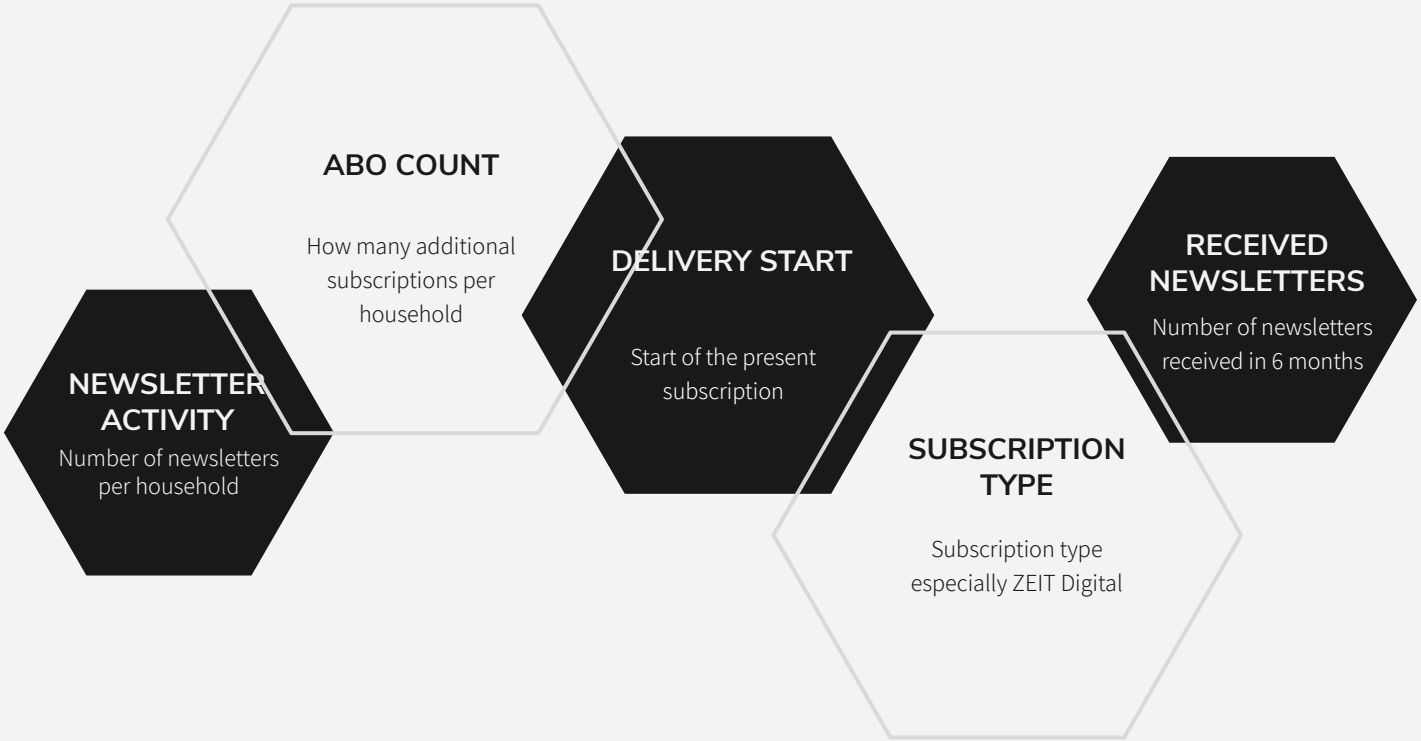


TOP FEATURES

17.6%
DELIVERY START

12.6%
NEWSLETTER ACTIVITY

TOP FEATURES



04

MACHINE LEARNING MODELS

WHAT IS OUR GOAL?

Emphasis on precision, because false alarms may cost money

PRECISION

Ratio between correctly predicted churns and false alarms

ACCURACY

How often has the model been right?

RECALL

How many churns are detected ?

F-BETA-SCORE

Weighted harmonic mean of precision and recall (Beta = 0.5)

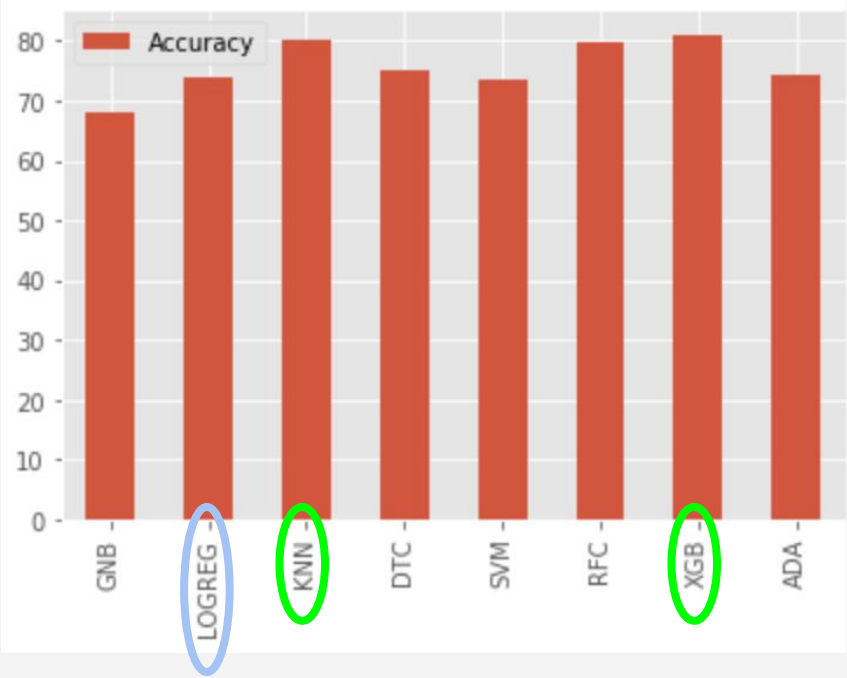
MACHINE LEARNING MODELS

USED MODELS:

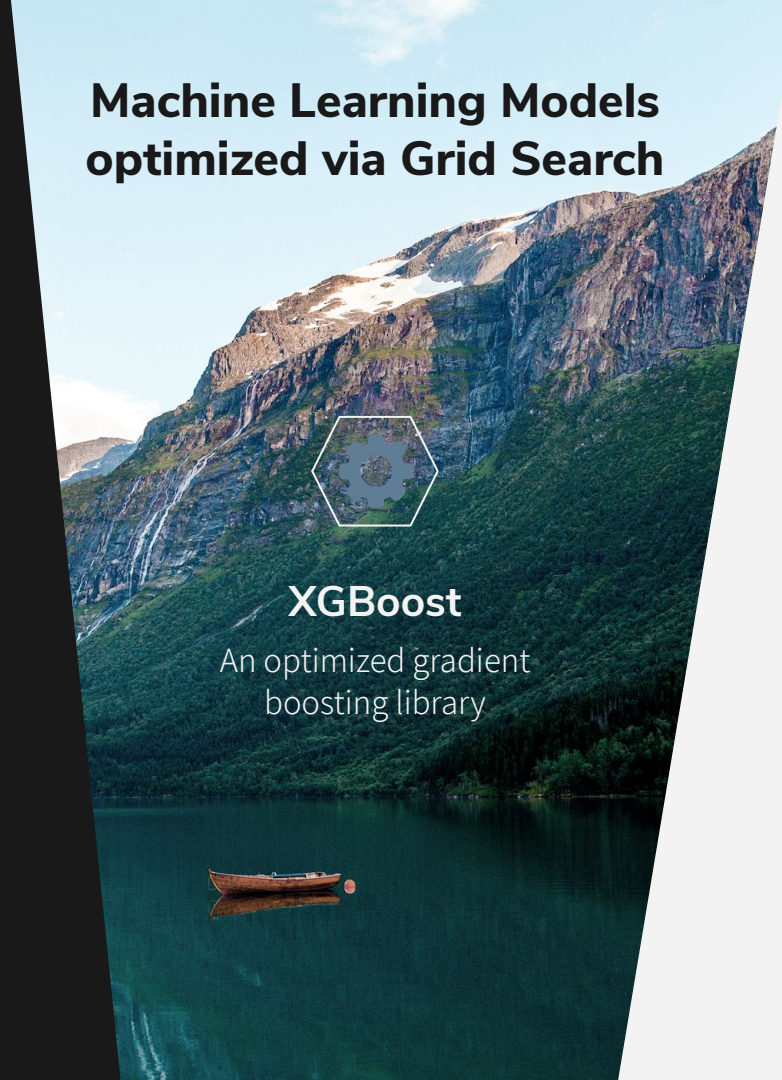
Baseline Model

Best Models

1. Gaussian Naive Bayes
2. Logistic Regression
3. K Nearest Neighbors
4. Decision Tree
5. Support Vector Machines
6. Random Forest
7. XGBoost
8. AdaBoost



Machine Learning Models optimized via Grid Search



K Nearest Neighbors

KNN classifies based on similarity with nearby data points

XGBoost

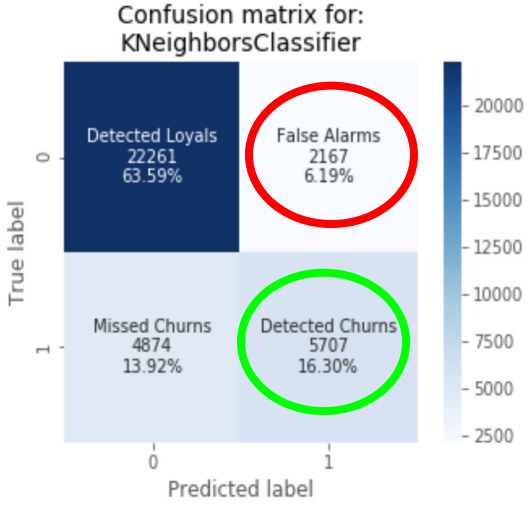
An optimized gradient boosting library

Random Forest

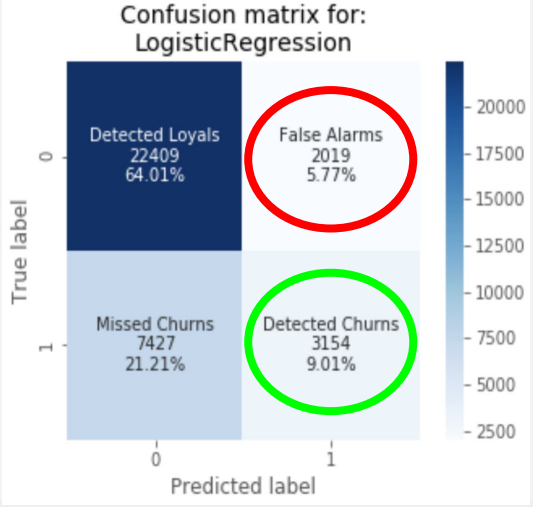
Randomly creates and merges multiple decision trees

BEST MODELS VS BASELINE MODEL

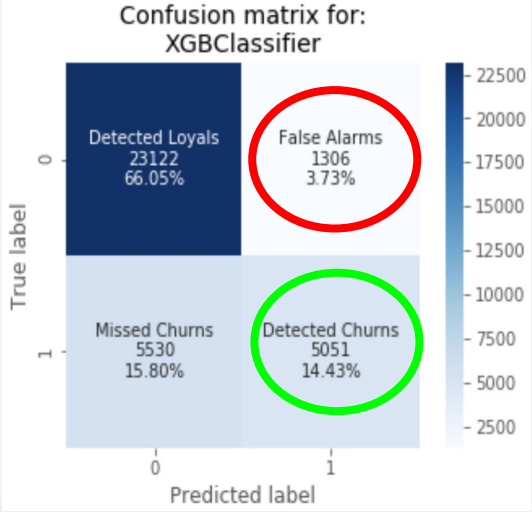
KNN



LOGISTIC REGRESSION BASELINE MODEL



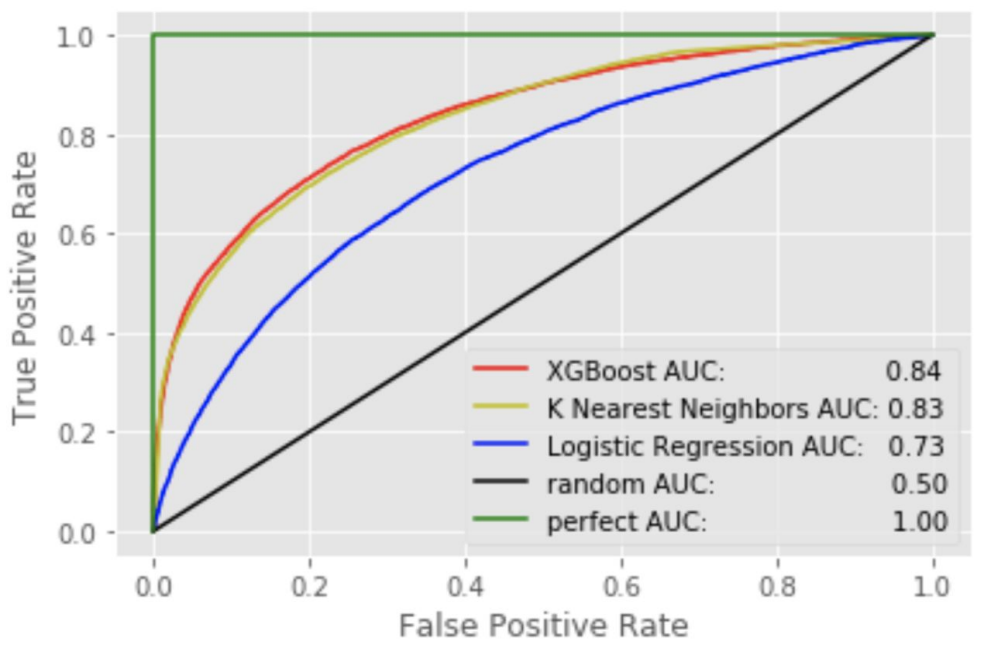
XGBoost



ACCURACY:	**79.9%**	**73.0%**	**80.5%**
RECALL:	**53.9%**	**29.8%**	**47.7%**
PRECISION:	**72.5%**	**61.0%**	**79.5%**

COMPARISON OF ROC AUC SCORE

- TRUE POSITIVE RATE = RECALL
- FALSE POSITIVE RATE = FALSE ALARMS / ALL LOYALS

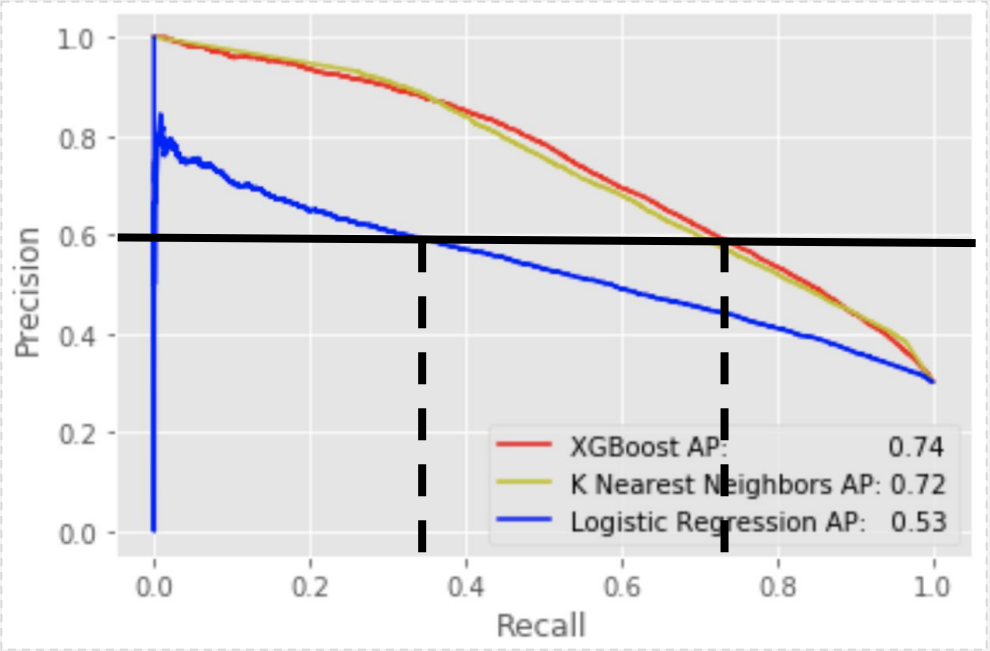


PRECISION RECALL CURVE

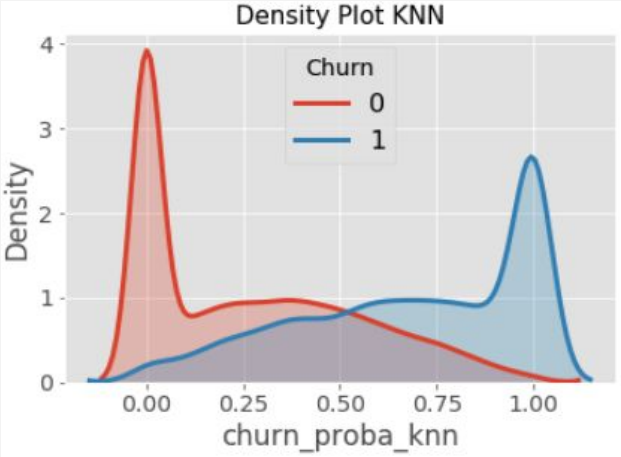
Choosing Precision 60%:

- Logistic Regression achieves **33% Recall**
- XGBoost / KNN achieve **73% Recall**

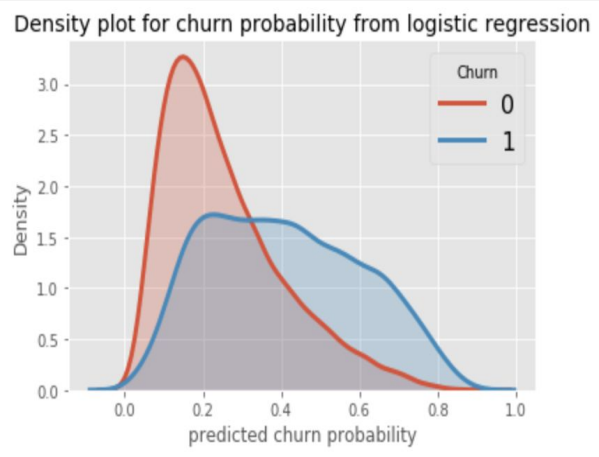
→ more than twice as many detected churns at the same rate of false alarms



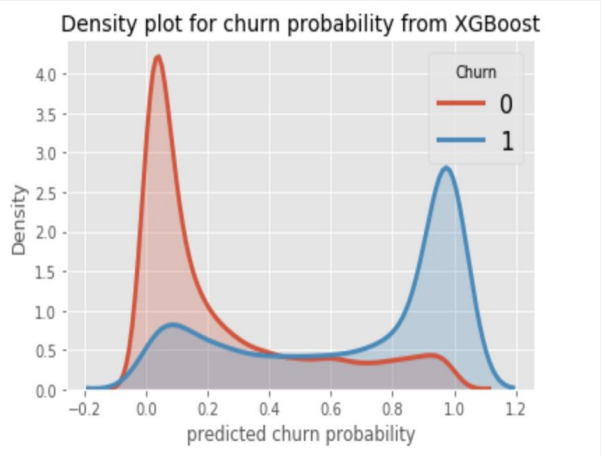
SELECTIVITY OF THE MODELS



KNN with balancing



Base model (Logistic Regression)



XGBoost without balancing

SUMMARY

With judicious feature selection and engineering we were able to tune several machine learning models such that they outperform the logistic regression model.

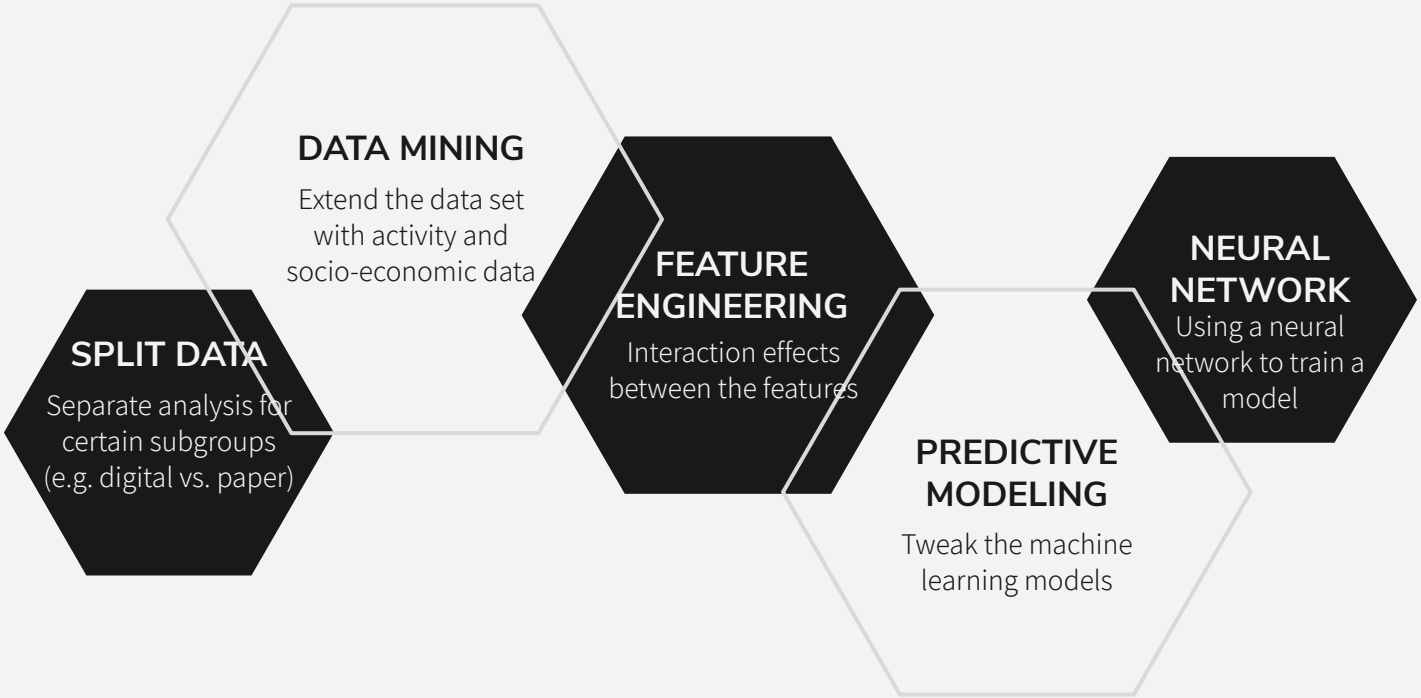
This model improvement enables publishers to predict churn much more reliably than before.

Financial resources for customer retention can thus be targeted more precisely.

05

FUTURE WORK

FUTURE WORK


DATA MINING
Extend the data set with activity and socio-economic data
SPLIT DATA
Separate analysis for certain subgroups (e.g. digital vs. paper)
FEATURE ENGINEERING
Interaction effects between the features
NEURAL NETWORK
Using a neural network to train a model
PREDICTIVE MODELING
Tweak the machine learning models

THANKS

Does anyone have any questions?

fuhrer_sebastian@web.de
https://github.com/booty7
Marco Zagermann
https://github.com/mz2357