“Zeit” Subscribers and Unsubscribers in the Light of Data Science

Marco Zagermann

Overview

1. What is churn prediction?
2. The dataset
3. Who is likely to churn?
4. Feature selection
5. ML models
6. Conclusion and outlook

1. What is churn prediction?

Churn prediction

A common problem of many newspapers and magazines:

Subscribers may end their subscription ("churn")

→ Negative effect on revenues

→ It is usually easier to prevent churn than attracting new customers

→ But: This requires that one knows beforehand who is likely to churn soon

→ Churn prediction!

2. The dataset

The original dataset

- **209 000** subscribers of **“Die Zeit”** (on paper and/or digital)
- **171** features
- Only subscriptions that were **still active in May 2019**
- **Starting dates** of those subscription: **2013 - 2019**
- Subscription cancellations **(“churns”)** from **June 2019 to May 2020**

The overall “churn probability” in the dataset: 30.2 %

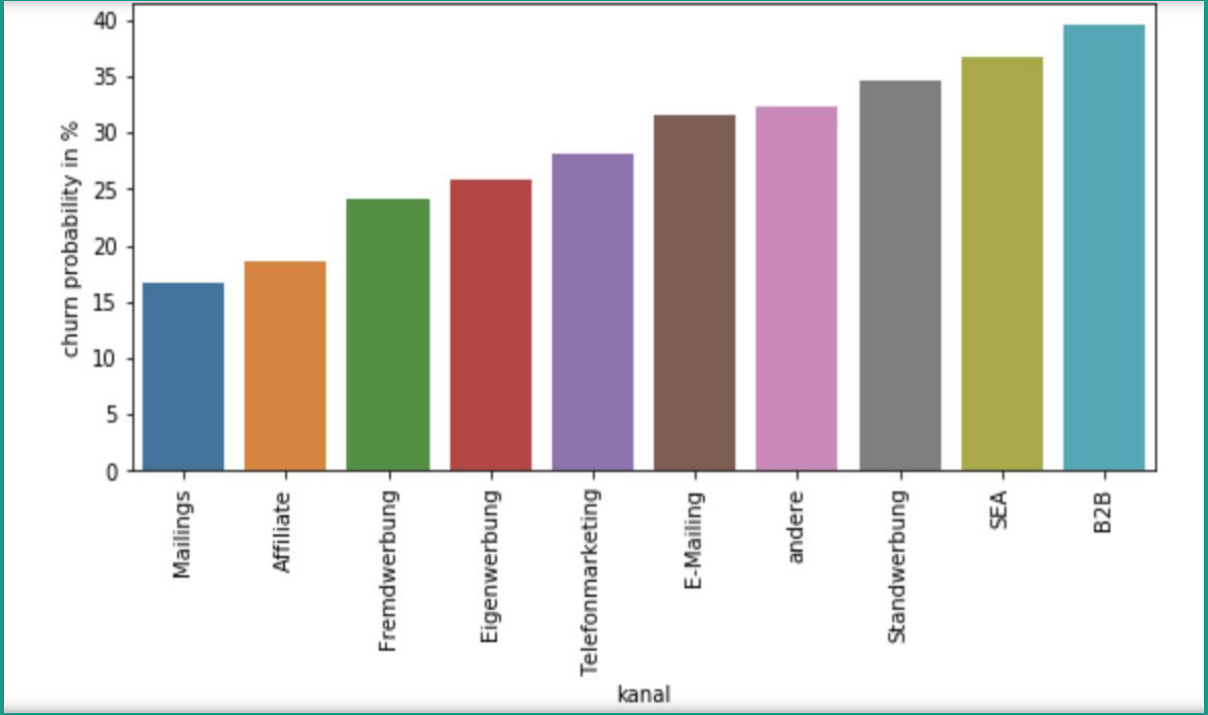
Goal:

Predict which subscribers are most likely to churn in the near future!

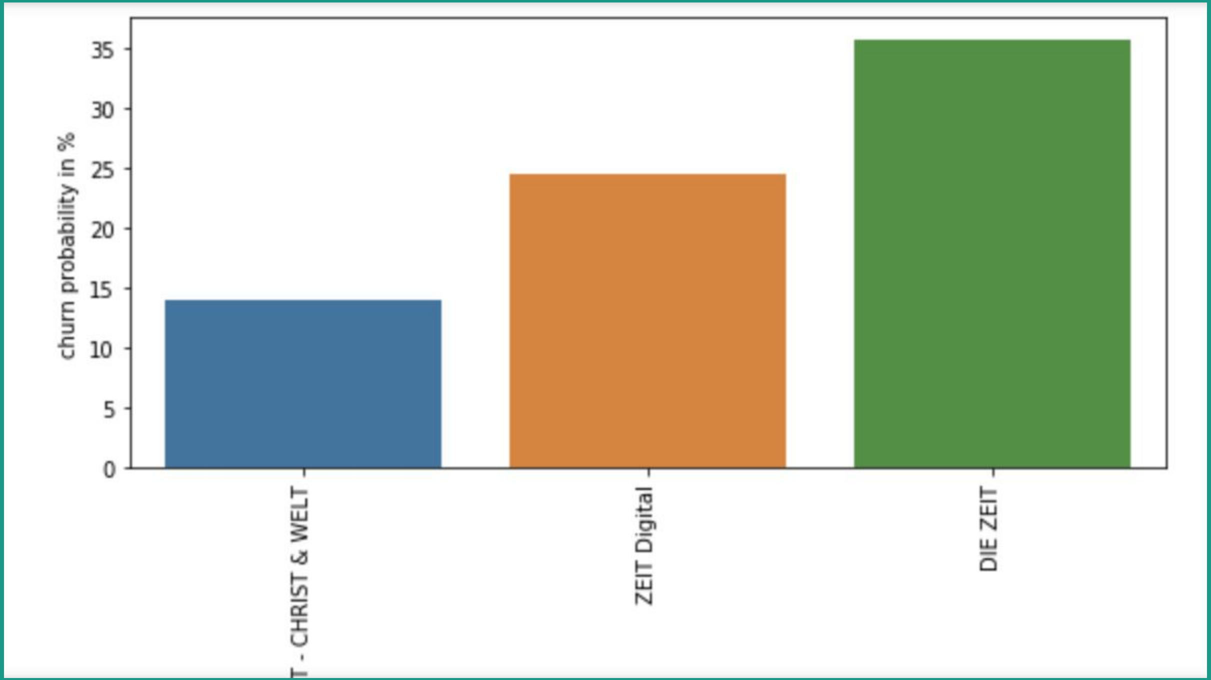
3. Who is likely to churn?

Churn probability by subgroups

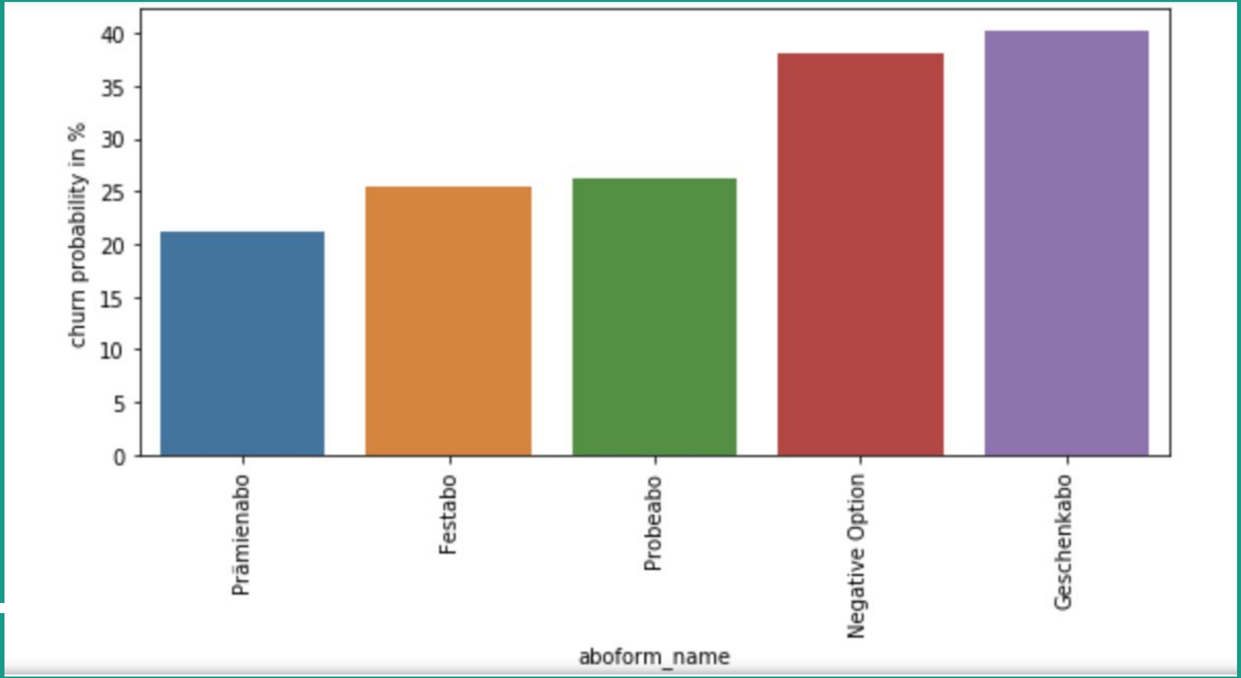
Channel of recruitment:



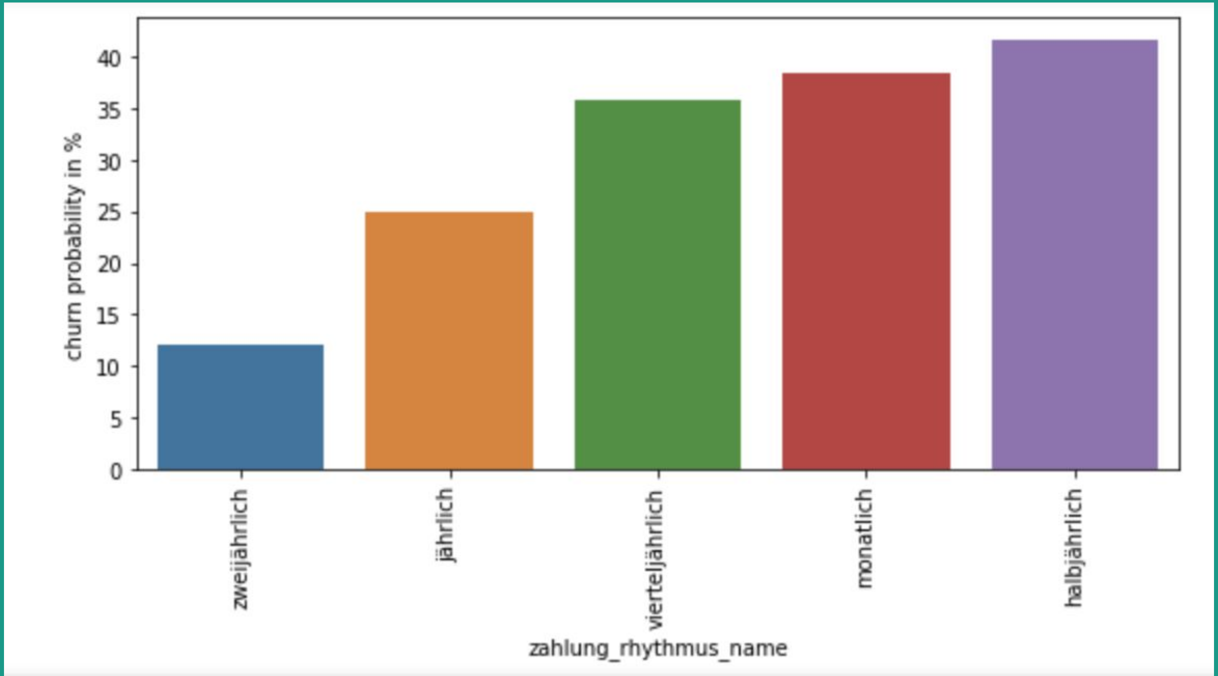
Digital vs. paper vs. Christ & Welt:



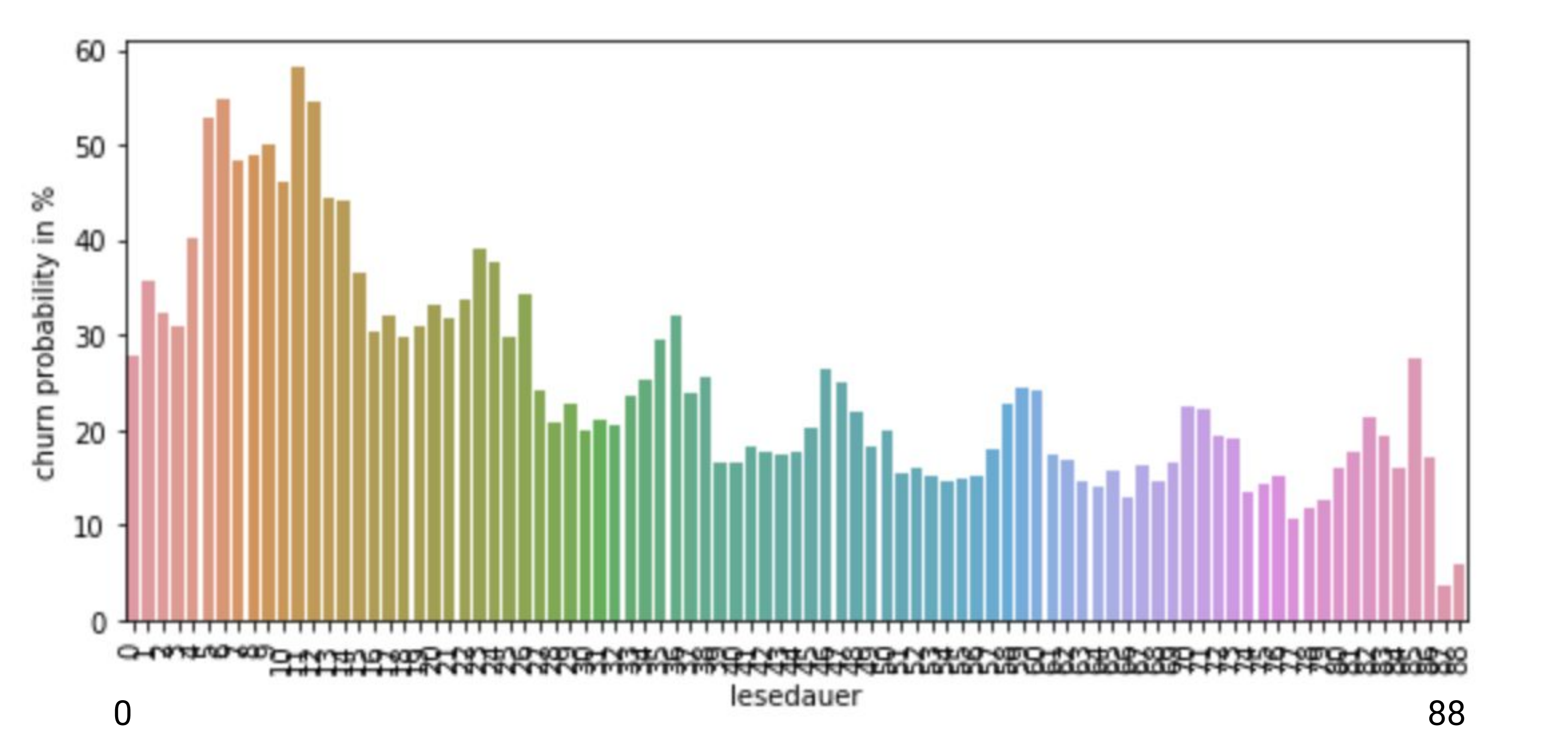
Type of subscription:



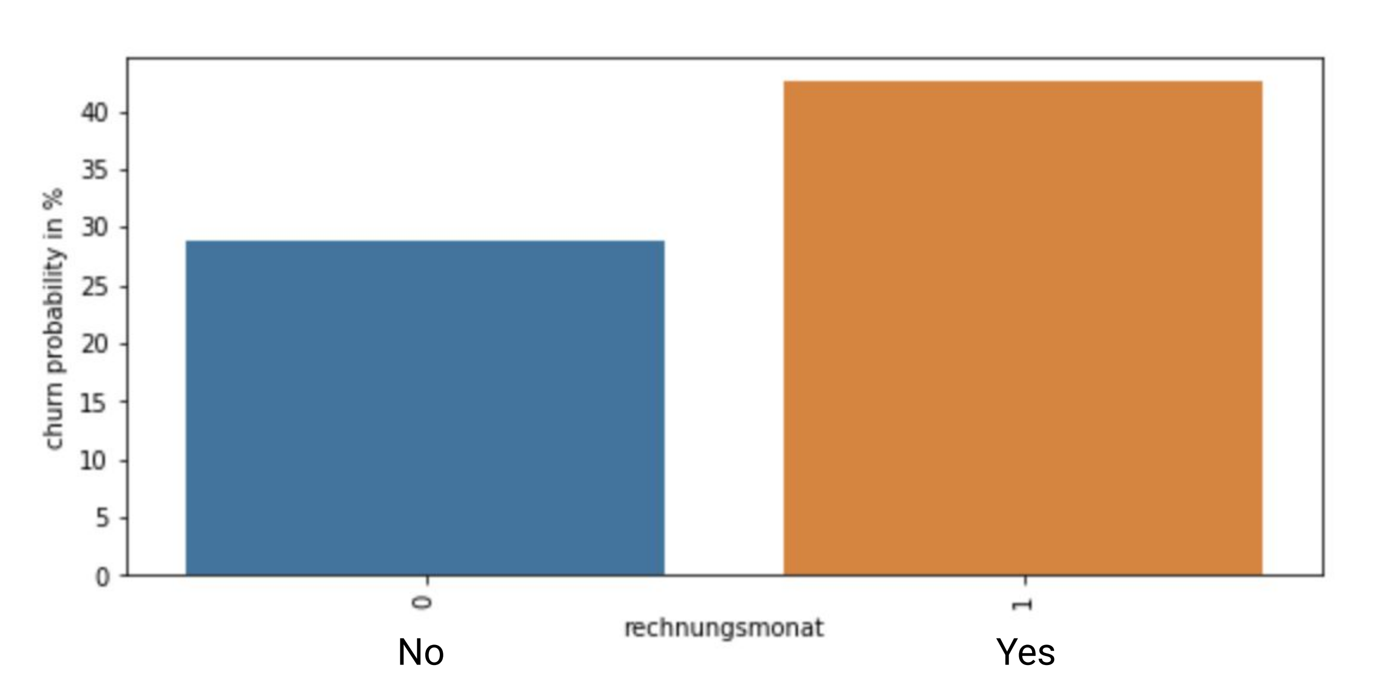
Rhythm of payment:



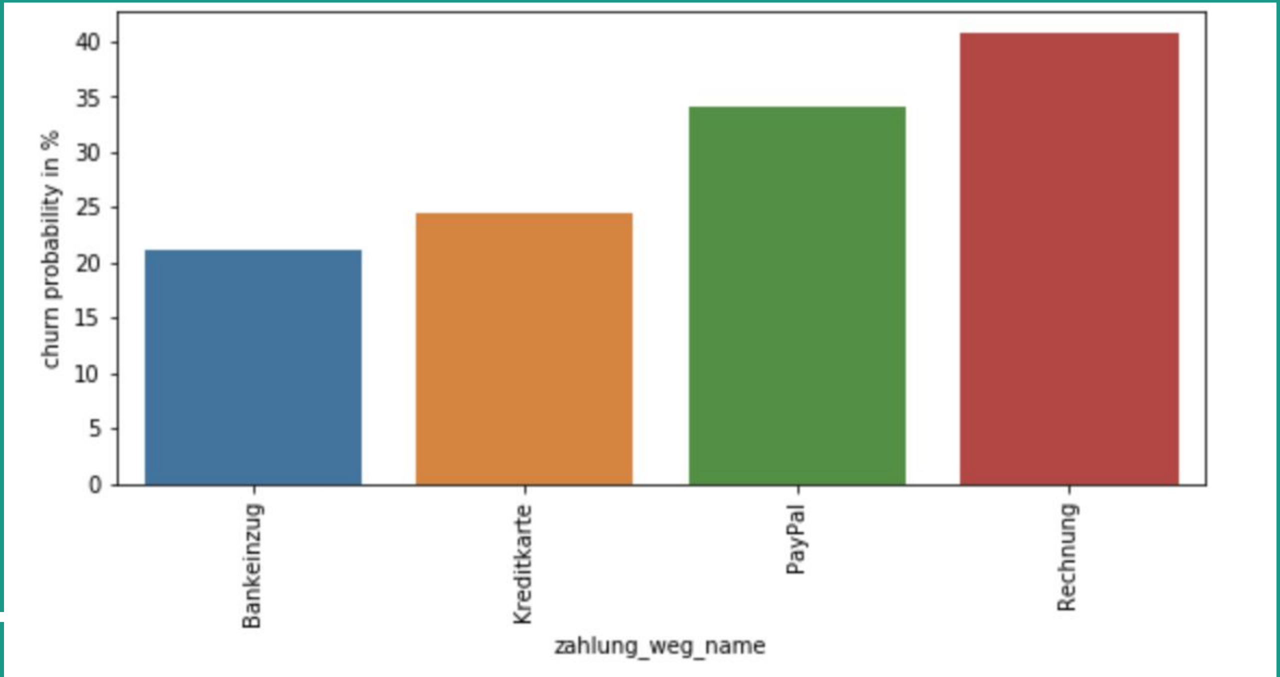
Months of reading:



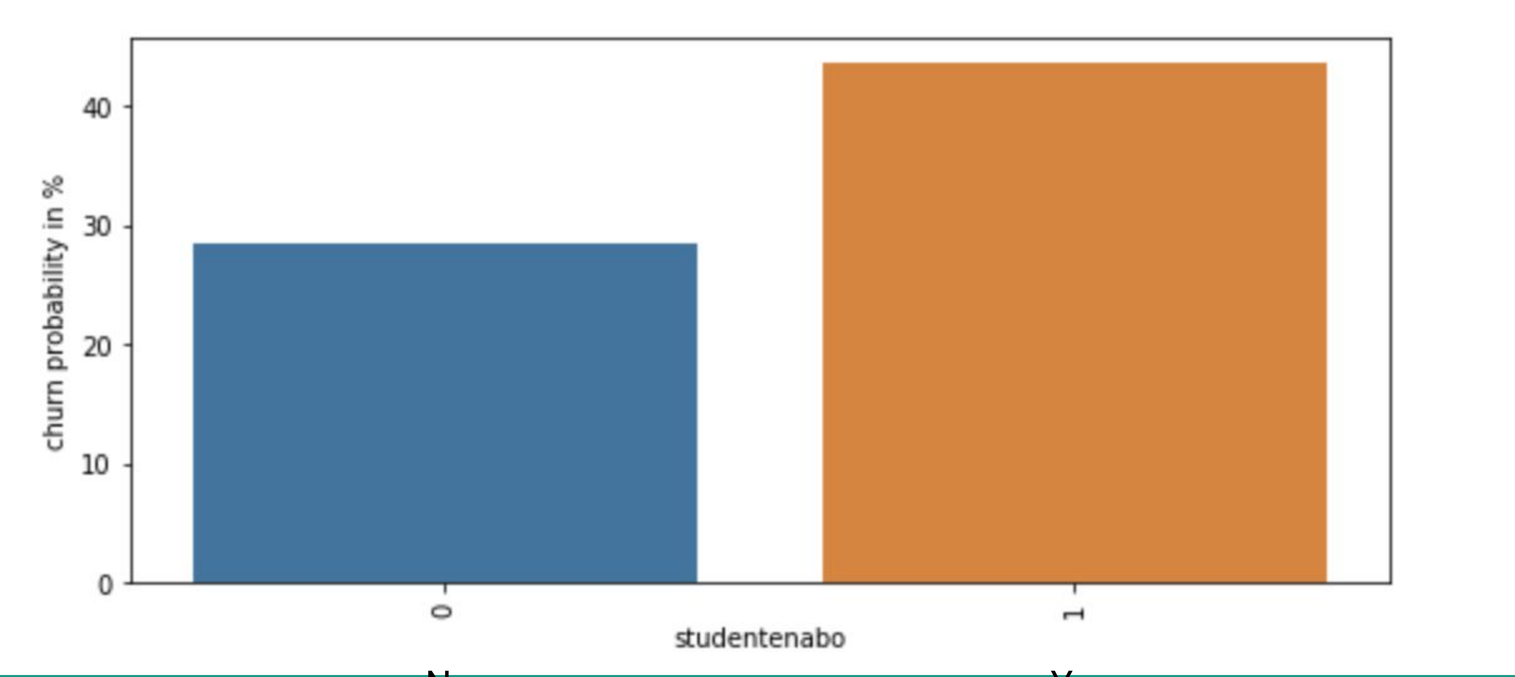
Billing month?:



Method of payment:

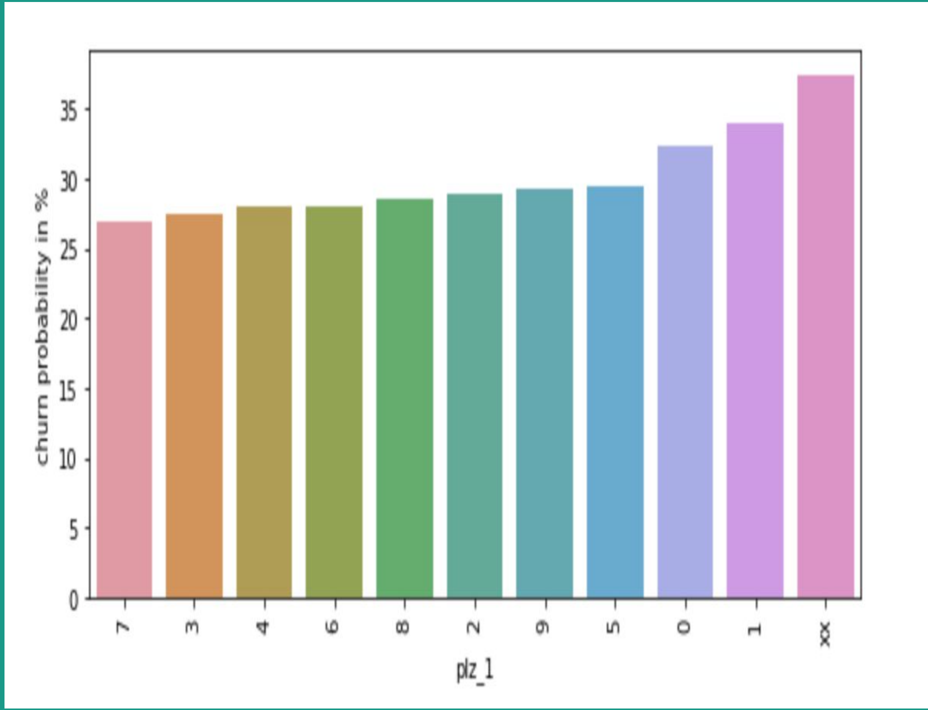


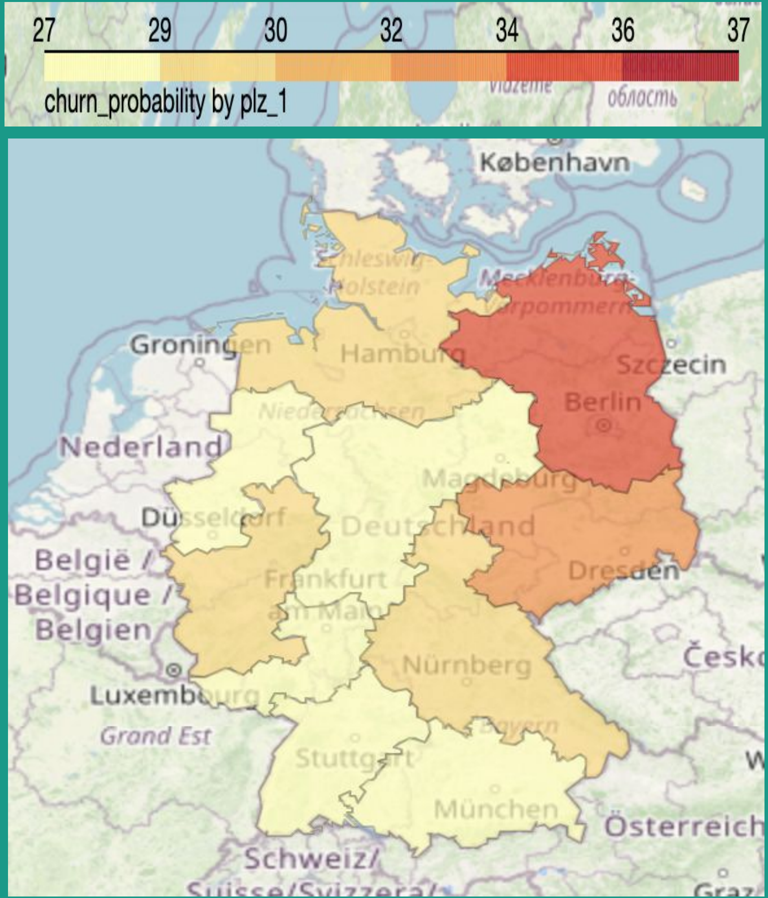
Student subscription?:



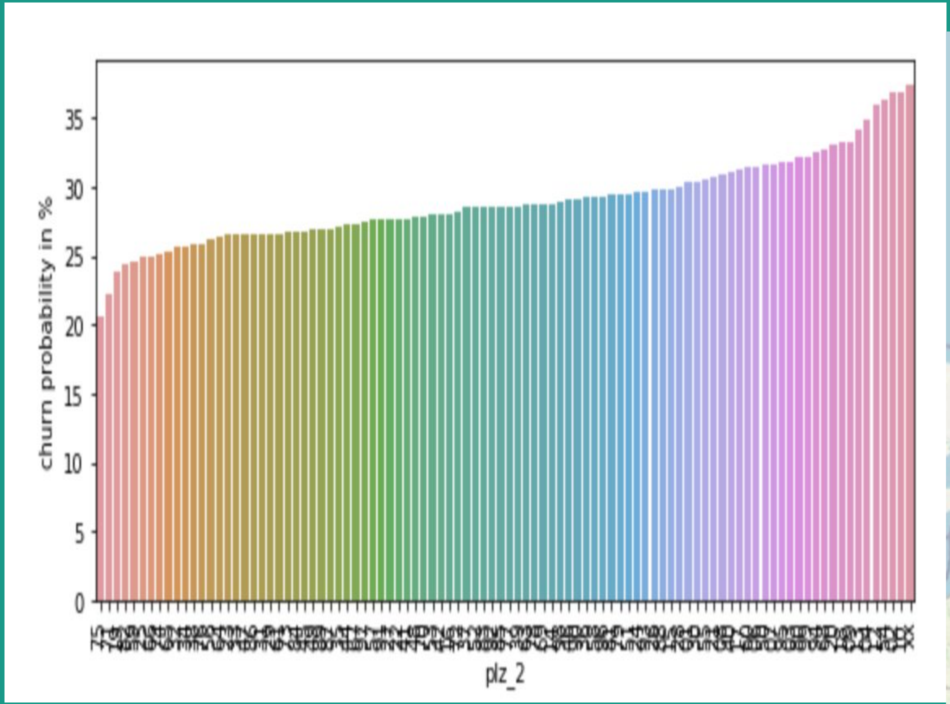
No Yes

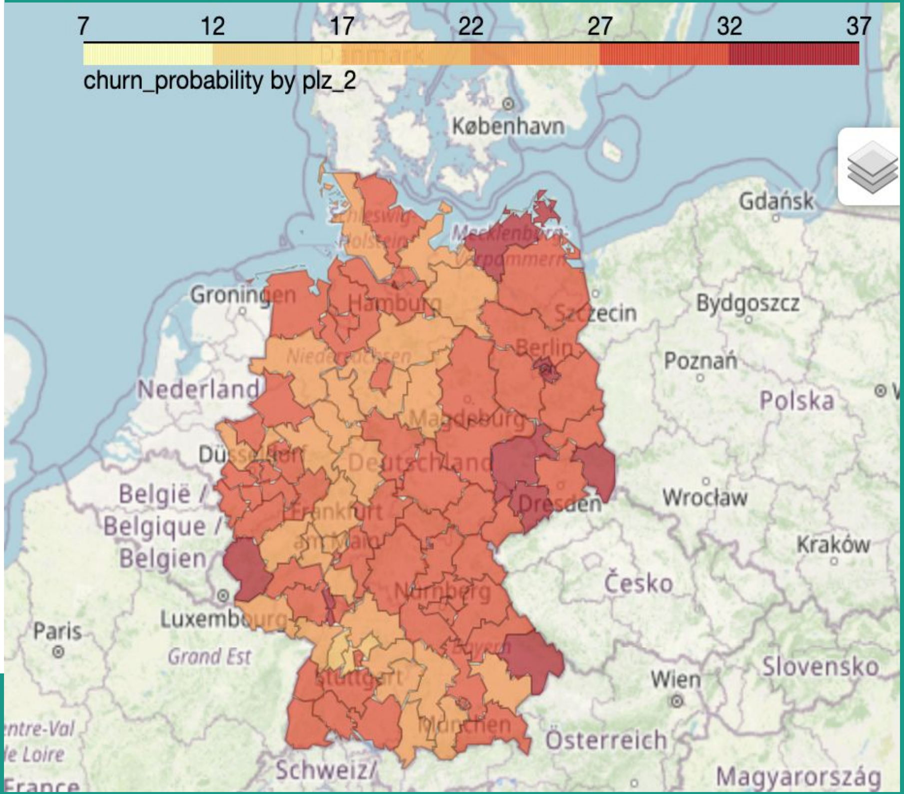
Postal code (one digit):





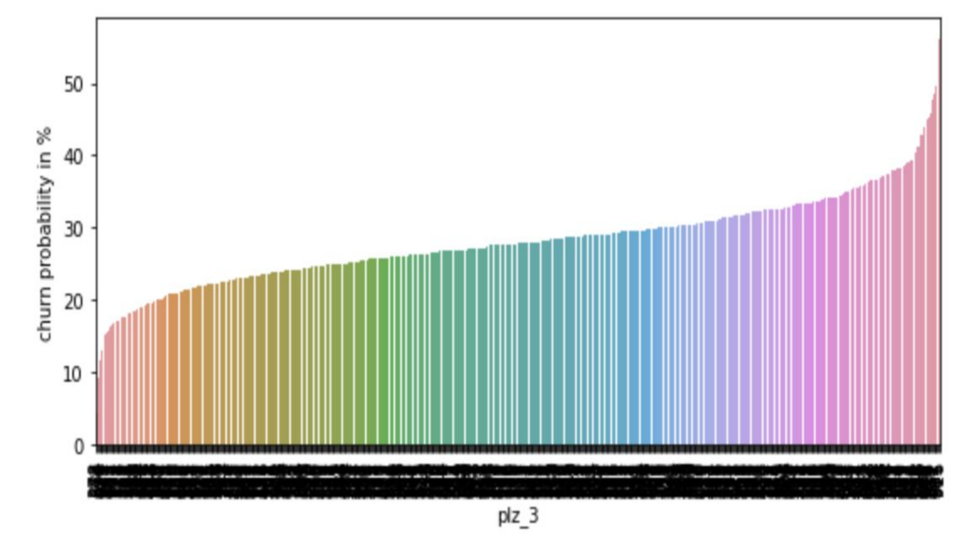
Postal code (two digits):

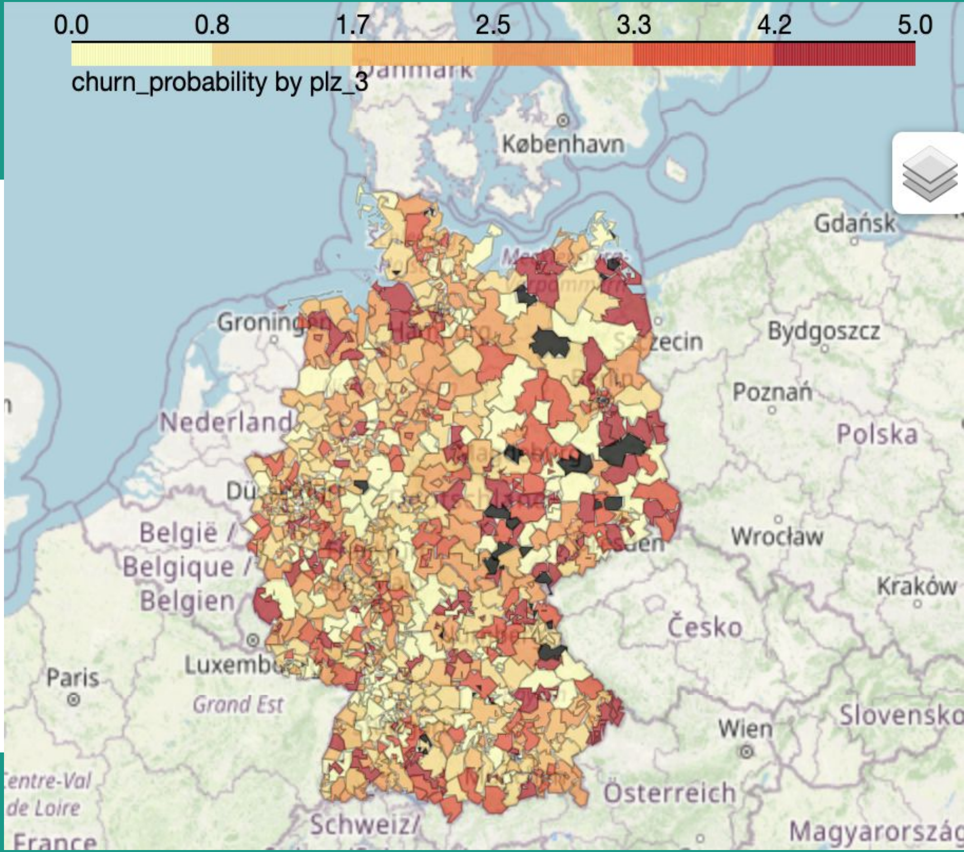




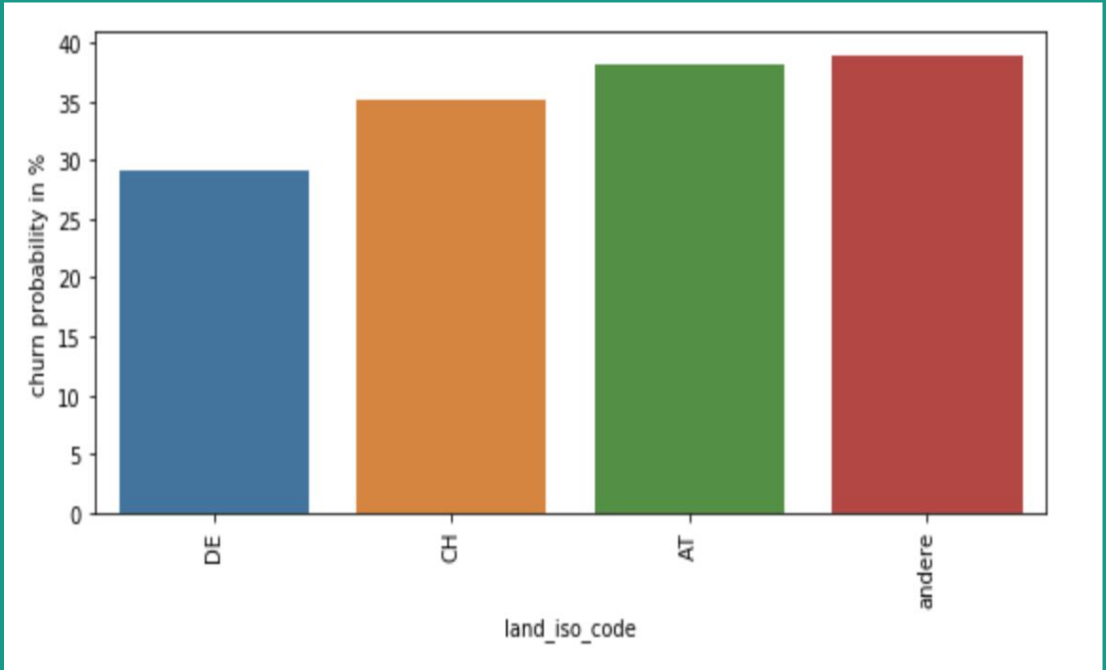
Postal code (three digits):

Quantiles are plotted here:

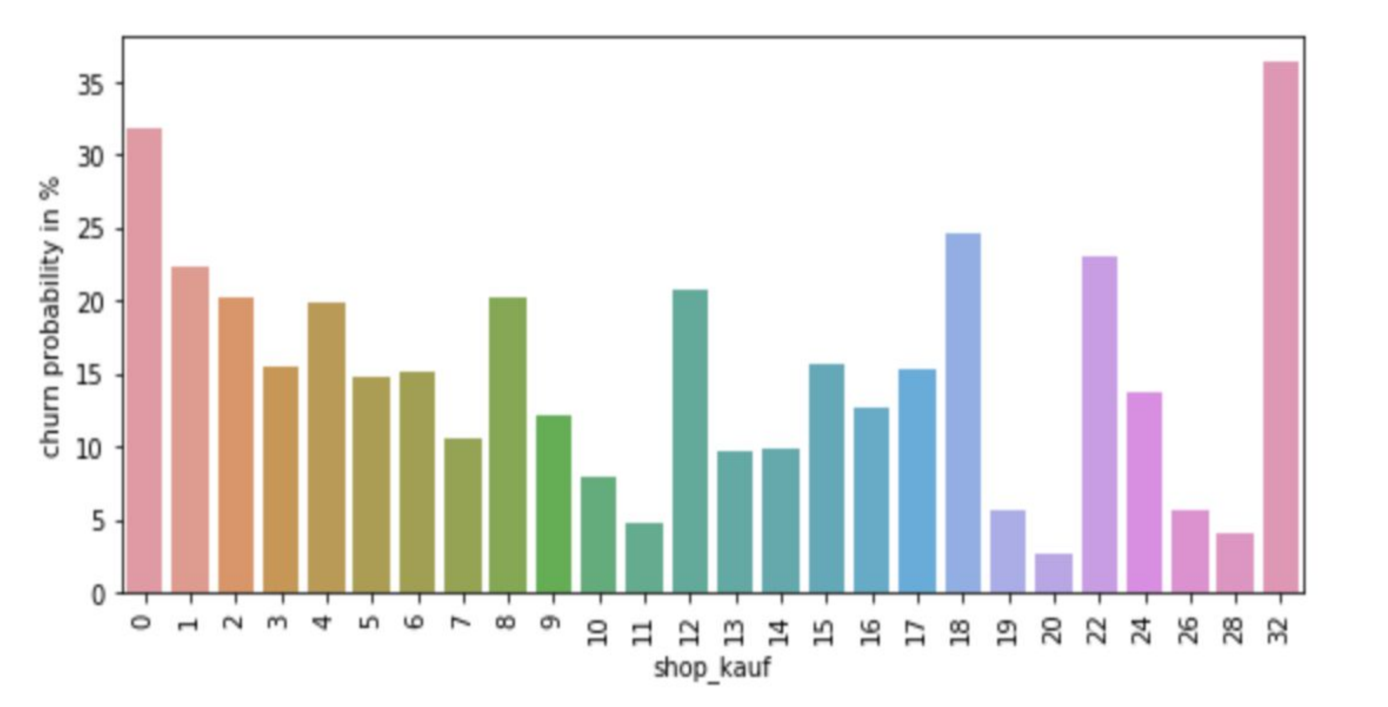




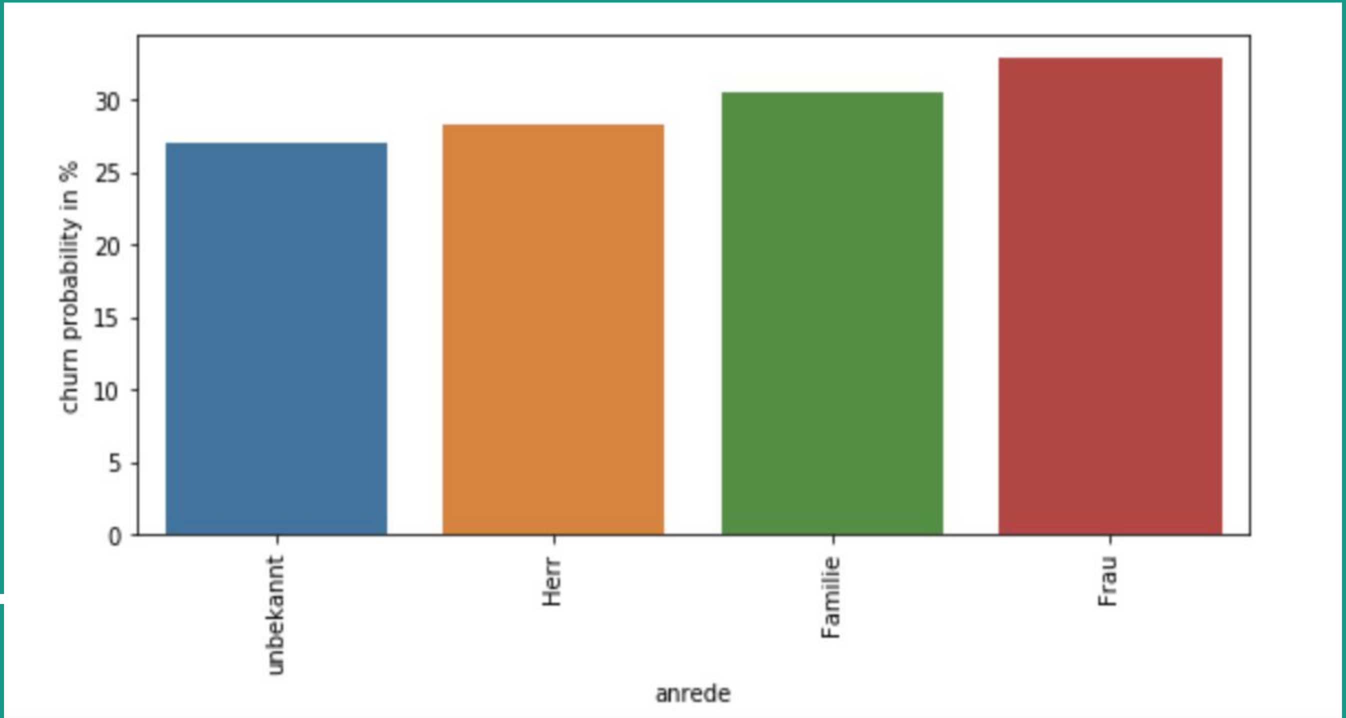
Country of residence:



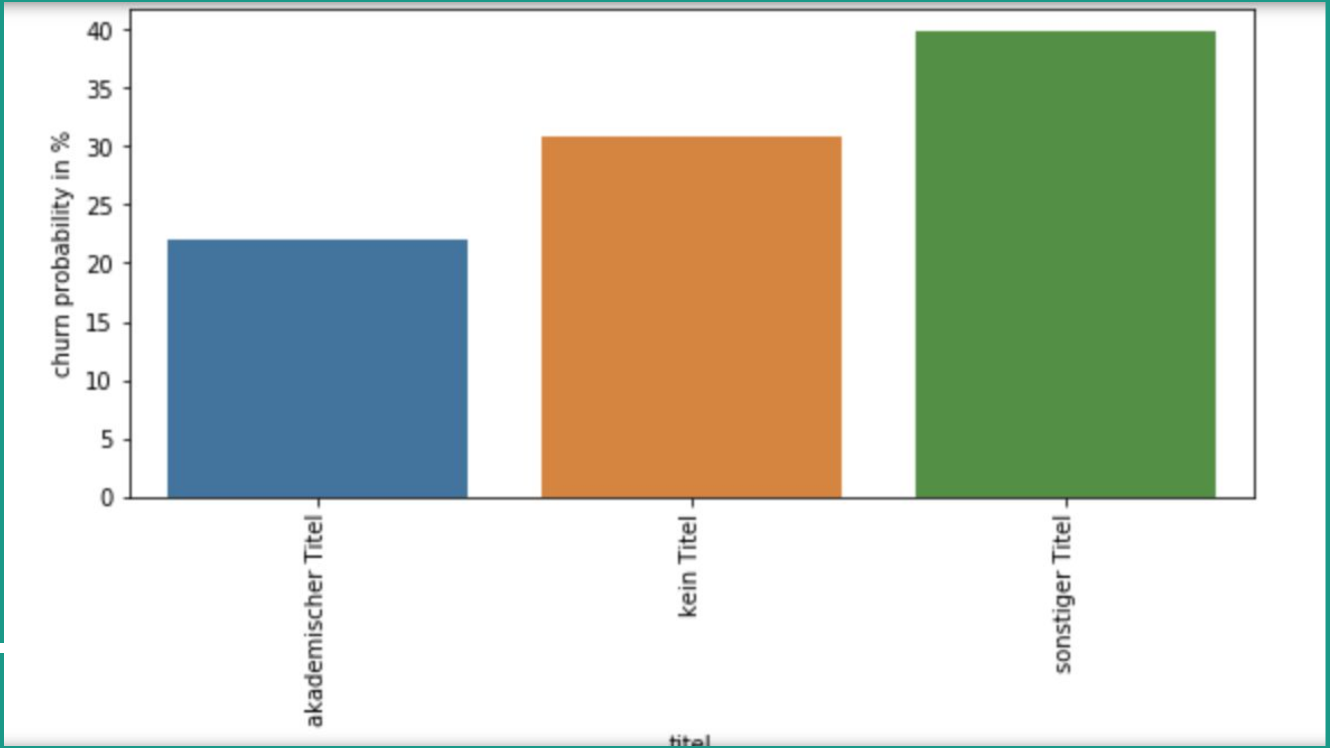
Shop purchases:



Mr./Mrs./Family:



Title:



churn probability in %
40
35
30
25
20
15
10
5
0
akademischer Titel
kein Titel
sonstiger Titel
titel

4. Feature selection

Two problems

Problem 1:

- **Dataset contains many categorical variables**
- Some of them (city of residence, postal code etc.) take on many values
- Naive treatment would lead to 11 000 dummy variables

Problem 2:

- **Many subgroups with high or low churn probability have a rather small size**
- Limits the predictivity of the corresponding feature

→ Feature selection!

Three methods

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

→ For tuning a classifier choose that feature set which works best with that classifier

→ Feature selection!

5. ML models

Models used

1. Gaussian Naive Bayes
2. Logistic regression
3. K nearest neighbors
4. Decision trees
5. Support vector machines
6. Random forests
7. XGBoost
8. AdaBost

→ Grid search

Randomized search

The best models:

```
[[22445  1983]
 [ 5568  5013]]
Accuracy: 0.784312605330058
Precision: 0.716552156089194
Recall: 0.473773745392685
ROC_AUC: 0.696298203955532
AP: 0.49852849138288413
f1: 0.5704045058883769
fbeta: 0.6499416569428238
```

Random forest optimized for the fbeta score

```
[[22435  1993]
 [ 6155  4426]]
Accuracy: 0.767259847467794
Precision: 0.6895155008568313
Recall: 0.4182969473584727
ROC_AUC: 0.6683551217879641
AP: 0.4642341632385613
f1: 0.5207058823529412
fbeta: 0.610364895054748
```

XGBoost optimized for accuracy

The best models:

```
[[22951  1477]
 [ 6718  3863]]
Accuracy: 0.7659173355422891
Precision: 0.7234082397003745
Recall: 0.36508836593894717
ROC_AUC: 0.6523124816431268
AP: 0.4560014452356122
f1: 0.4852710256893411
fbeta: 0.6047086816317585
```

XGBoost optimized for the fbeta score

```
[[22270  2158]
 [ 6221  4360]]
Accuracy: 0.7606615441743552
Precision: 0.6689168456581773
Recall: 0.41205935166808433
ROC_AUC: 0.6618590519597995
AP: 0.45333060532827485
f1: 0.5099713433534124
fbeta: 0.5947671404796334
```

K nearest neighbors optimized for accuracy and the fbeta score

6. Conclusion and outlook

Summary

With judicious feature selection and tuning and selecting ML models, we are able to predict churn of "Zeit" subscribers with almost 78% accuracy and almost 72% precision at 47% recall.

Future work

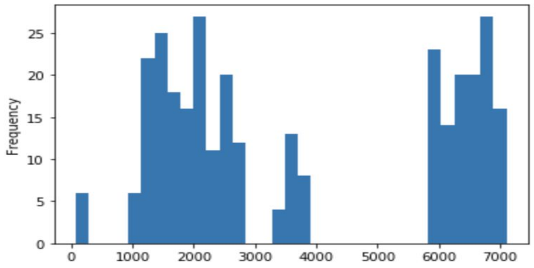
- There is a lot of unused information in the geographical features
  →Suitable aggregation, perhaps with external data
- Do some more feature engineering
- Use more ensemble methods
- Try a neural network
- Try some more balancing methods
- Analyse the effects of the measures that have been taken to avoid churn.

Thank you!

The truncated dataset

Problem:

- **Original dataset** contains many subscribers with **very high numbers of subscriptions** (up to **7000**, may include different kinds of publications)
- Presumably larger companies/institutions

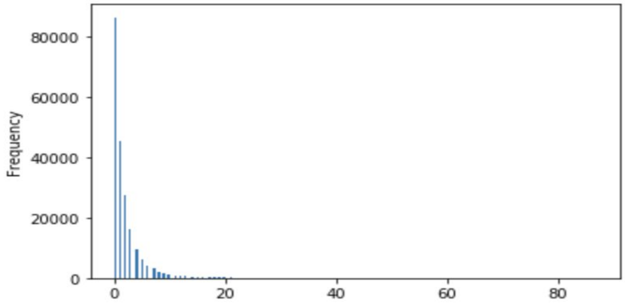


The extreme: Subscribers with >78 subscriptions
(Clusters around 1000 - 3000 and 6000 -7000 subscriptions)

The truncated dataset

Problem:

- **Original dataset** contains many subscribers with **very high numbers of subscriptions** (up to **7000**, may include different kinds of publications)
- Presumably larger companies/institutions



Exponential fall-off for < 20 subscriptions per subscriber

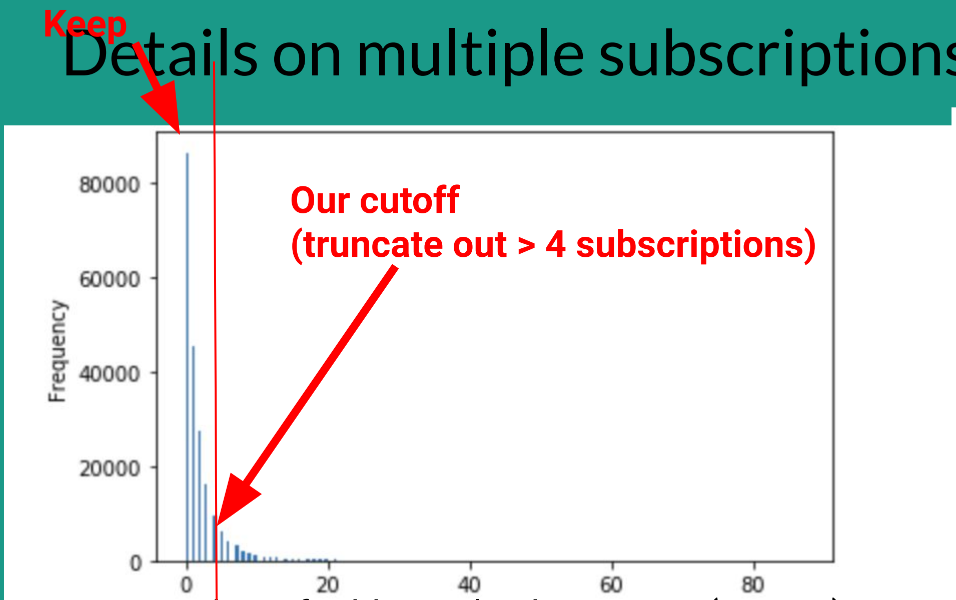
The truncated dataset

Problem:

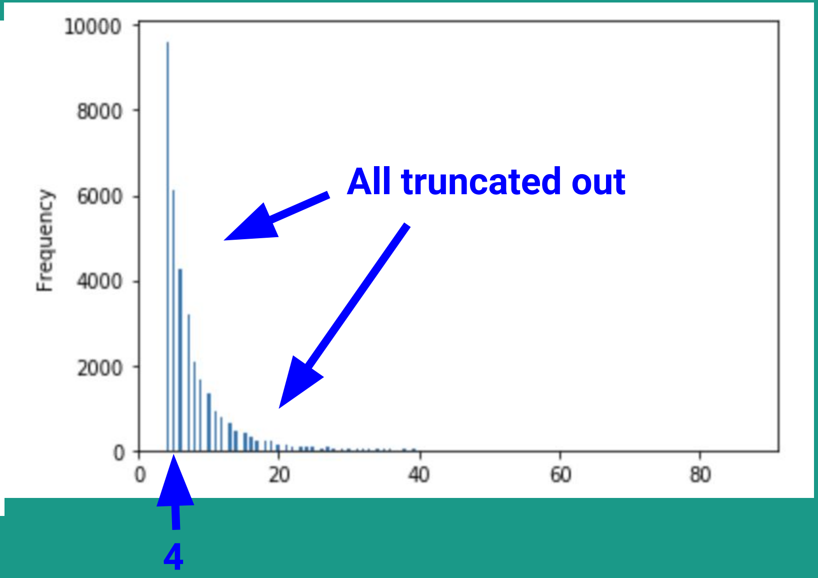
- **Original dataset** contains many subscribers with **very high numbers of subscriptions** (up to **7000**, may include different kinds of publications)
- Presumably larger companies/institutions
- **Hard to draw the line** to “ordinary” subscribers/housholds with given data
- **Our approach:** Truncate to subscribers with **at most four** subscriptions

Result: 209 000 → **175 000** (i.e. we drop 33 000 subscribers)

Details on multiple subscriptions



Number of additional subscription (min =0)
Total numbers of subscribers: 175 000



33 000 subscribers wiith **> 4 subscriptions (All truncated out)**

The target variable: “churn or not churn?”

- **Starting point:** 175 000 subscribers in **June 2019**
- 53 000 of them **cancel** their subscription in the reference period **June 2019-May 2020**
- **This gives the** overall “churn probability” **of** 30.2 %

Question:
Which of the 170 given features are **good predictors for a high churn probability?**

Groups of features

- Formal subscription features
- Subscription options
- Personal information
- Temporal features
- Location features
- Activity features

Formal subscription features:

kanal
objekt_name
aboform_name
zahlung_rhythmus_name
rechnungsmonat
zahlung_weg_name
studentenabo
unterbrechung

Subscription options:

zon_che,_opt_in
zon_sit_opt_in
zon_zp_grey
zon_premium
zon_boa
zon_kommentar
zon_sonstige
zon_zp_red
zon_app_sonstige
cnt_abo
cnt_abo_diezeit
cnt_abo_diezeit_digital
cnt_abo_magazin

cnt_umwandlungsstatus2_dkey
nl_zeitbrief
nl_zeitshop
nl_zeitverlag_hamburg

Personal information

anrede

titel

Temporal features:

lesedauer
liefer_beginn_evt
abo_registrierung_min
nl_registrierung_min

Location features

plz_1
plz_2
plz_3
ort
metropole
land_iso_code

Activity features

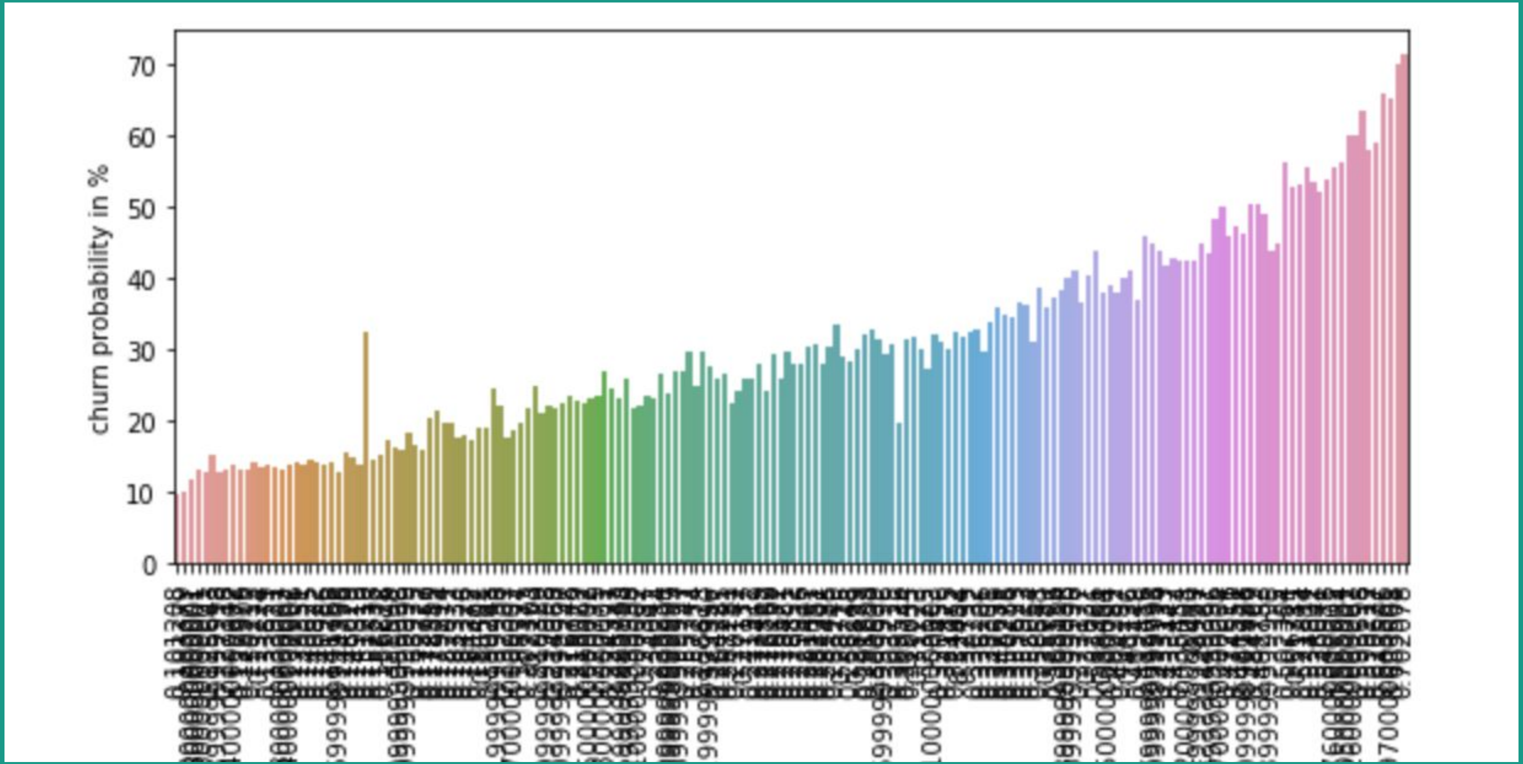
shop_kauf
email_am_kunden
nl_blacklist_sum
nl_bounced_sum
nl_aktivitaet
nl_sperrliste_sum
received_anzahl_1w
received_anzahl_1m
received_anzahl_3m
received_anzahl_6m
opened_anzahl_1w
opened_anzahl_1m

opened_anzahl_3m
openedanzahl_6m
clicked_anzahl_1w
clicked_anzahl_1m
clicked_anzahl_3m
clicked_anzahl_6m
unsubscribed_anzahl_1w
unsubscribed_anzahl_1m
unsubscribed_anzahl_3m
unsubscribed_anzahl_6m
openrate_1w
clickrate_1w

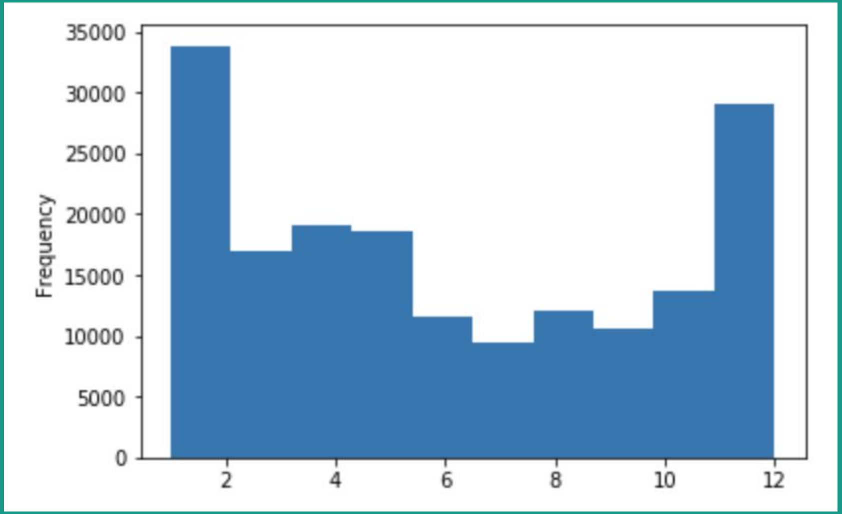
clickrate_1m
openrate_3m
clickrate_3m
received_anzahl_bestandskunden_1w
received_anzahl_bestandskunden_1m
received_anzahl_bestandskunden_3m
received_anzahl_bestandskunden_6m

\+ many more (altogether more than 100)

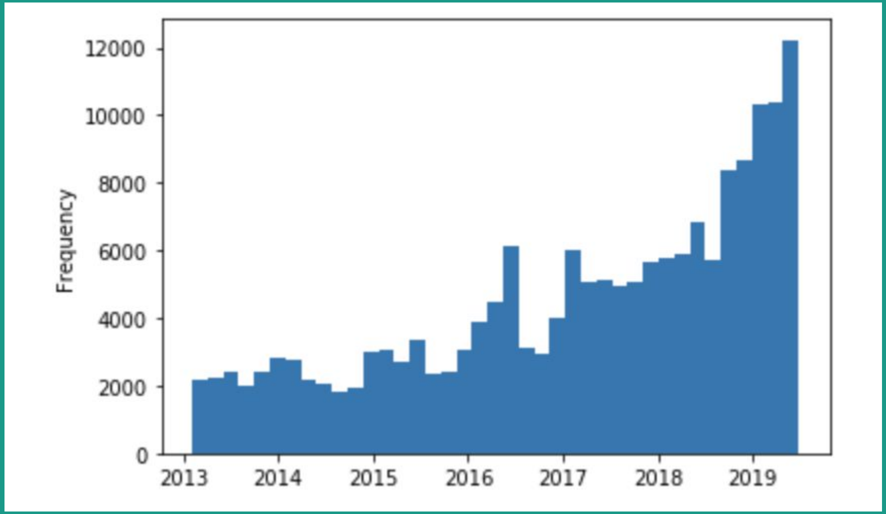
“Average churn” (based on months of reading and rhythm of payment):



Seasonal variation of begins of subscriptions



Average seasonal variation of begins of subscriptions



Temporal evolution of new subscriptions

Some open questions

- How should one best make use of the following location features?

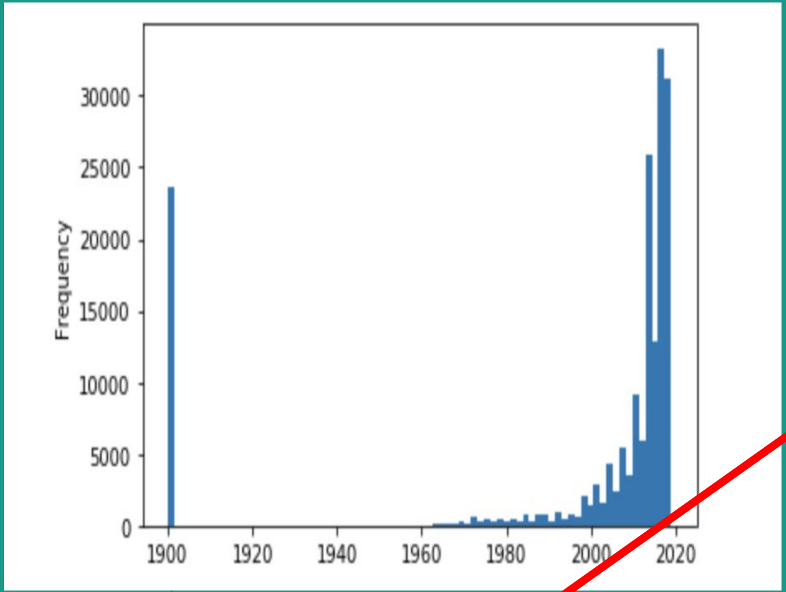
  plz_1 (eleven different values)
  plz_2 (97 different values)
  plz_3 (697 different values)
  ort (11 137 different values)

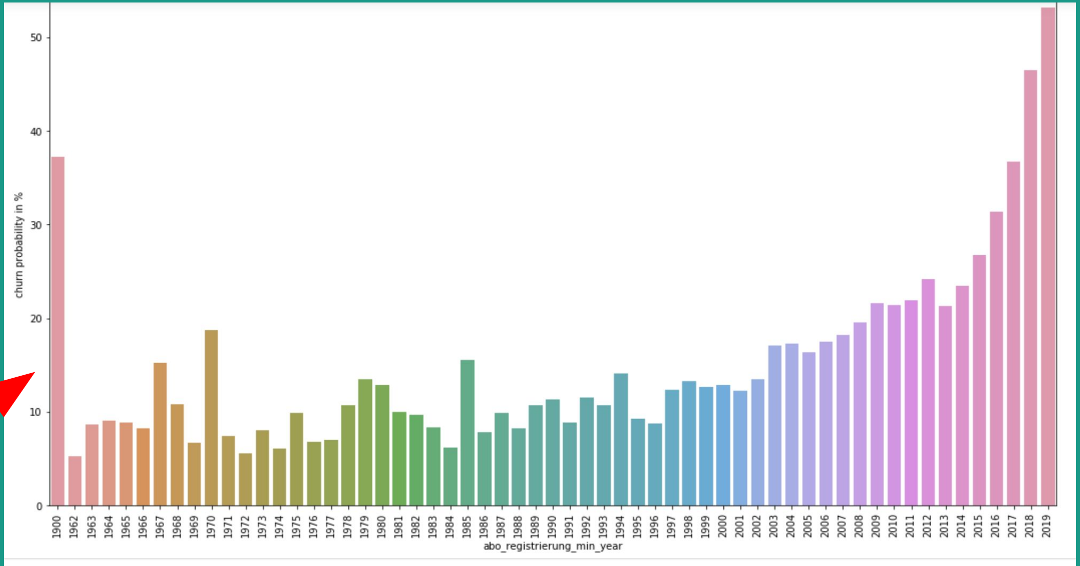
Some open questions

- How should one treat these two temporal features?

  abo_registrierung_min_year
  nl_registrierung_min_year

Earliest year of registration





Histogramm

Churn probability according to earliest year of registration

1900!

1900 is not a real date, but due to some operational cutoffs. Treating this feature as a numerical variable could thus lead to unwanted effects.

Some open questions

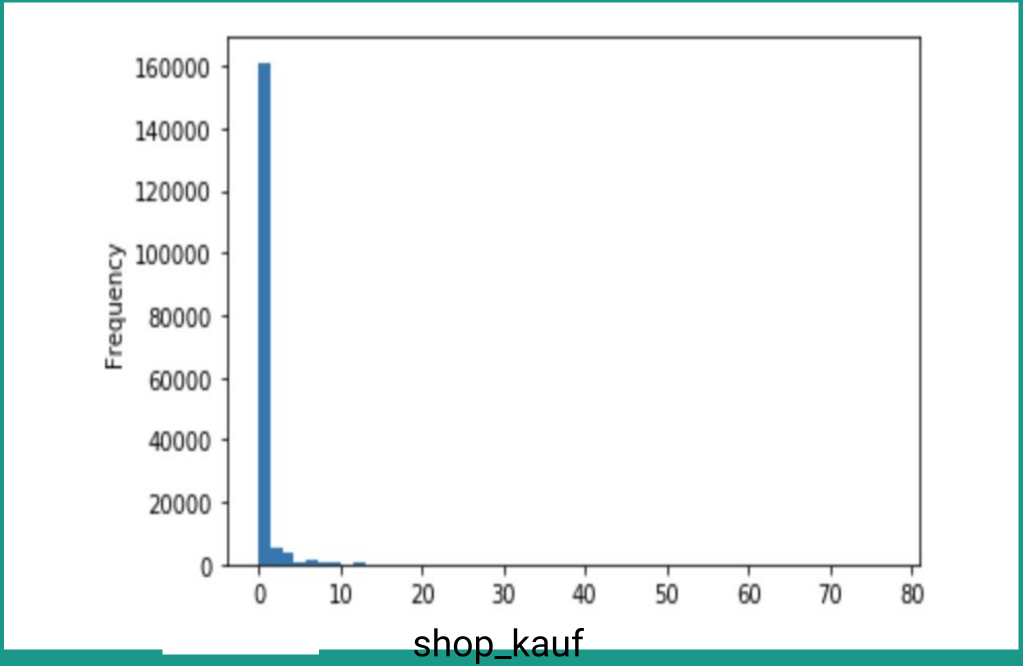
- How should one treat the very skewed features?

Two out of many examples:

shop_kauf
opened_anzahl_zeitbrief_3m

Shop purchases

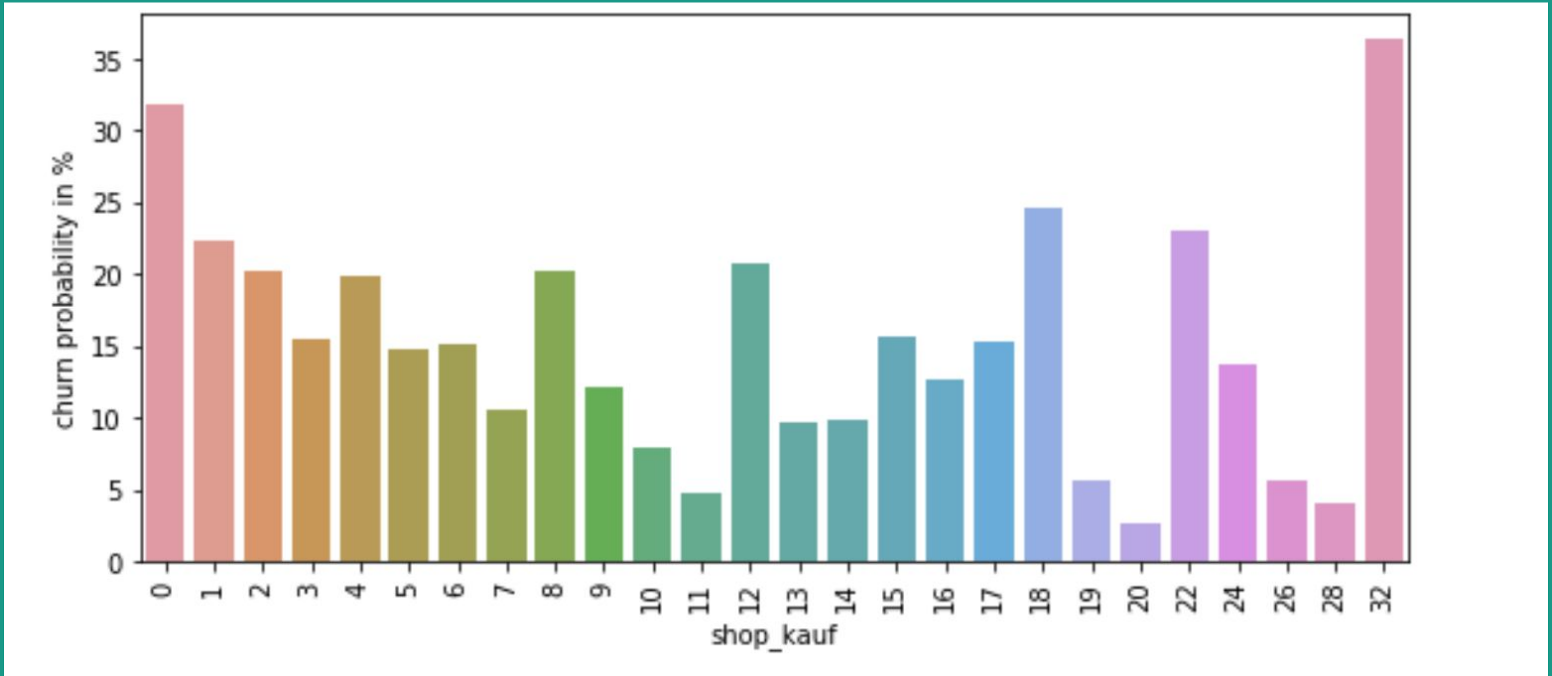
(n=175 000)



0	153802
1	7327
2	5188
3	2066
4	1985
6	1016
5	847
8	563
7	417
10	337
9	245
12	227
11	169
14	131
13	103
16	87
18	73
17	59
15	51
22	39
20	37
19	35
26	35

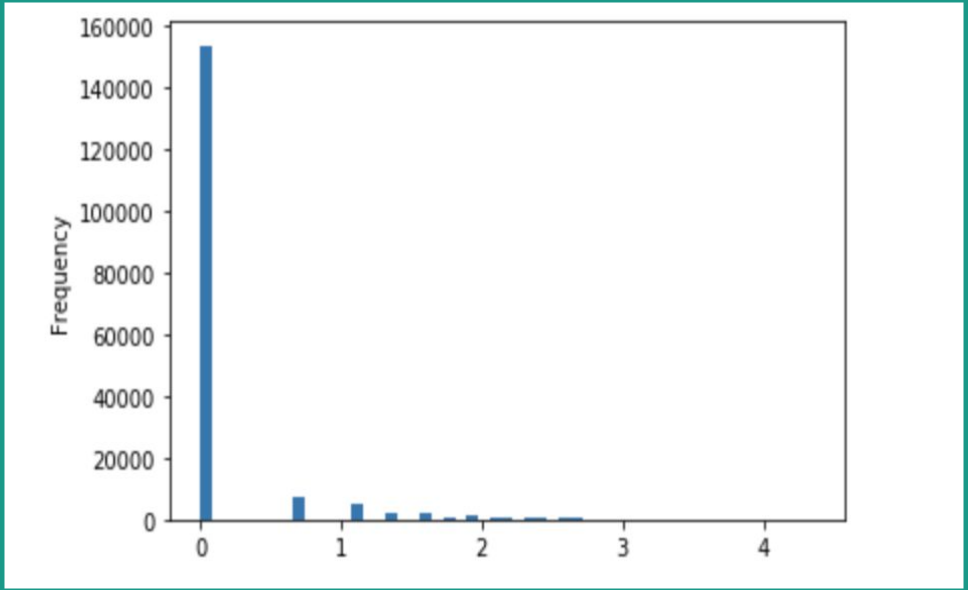
Not normally distributed and highly skewed!

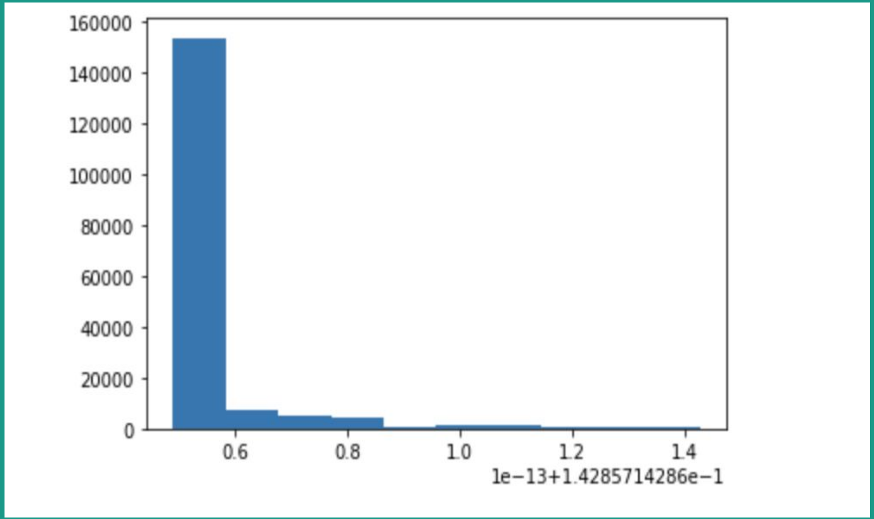
Shop purchases



Churn probability by shop_kauf: shop_kauf has an influence on churn
(The large fluctuations for the larger values are probably due to the small number of cases)

Shop purchases



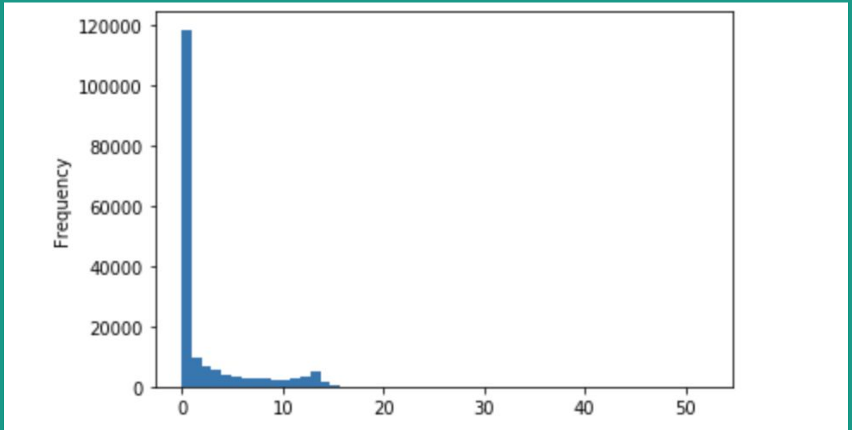


Distribution of logarithm of shop_kauf

boxcox of shop_kauf with parameter -7

Doesn't look very normally distributed

Opened number of “Zeitbrief” in three months



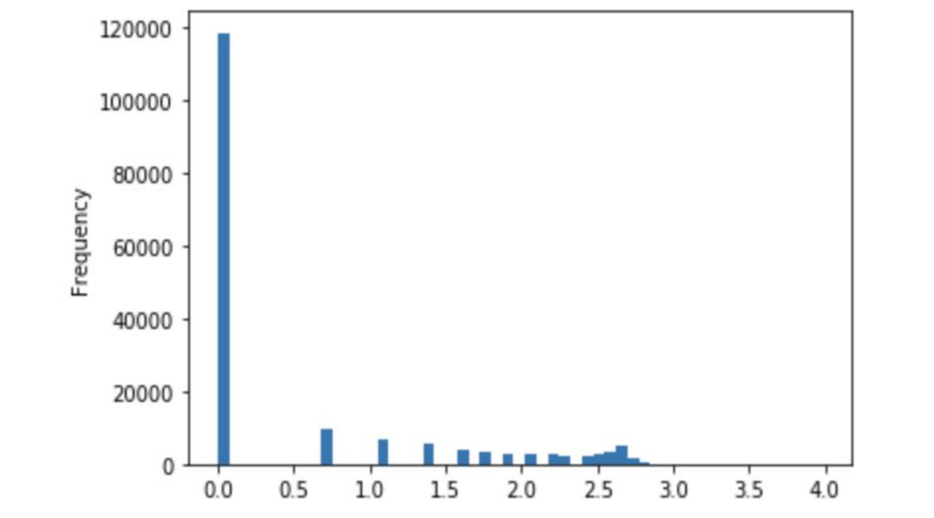
opened_anzahl_zeitbrief_3m

Not normally distributed and highly skewed!

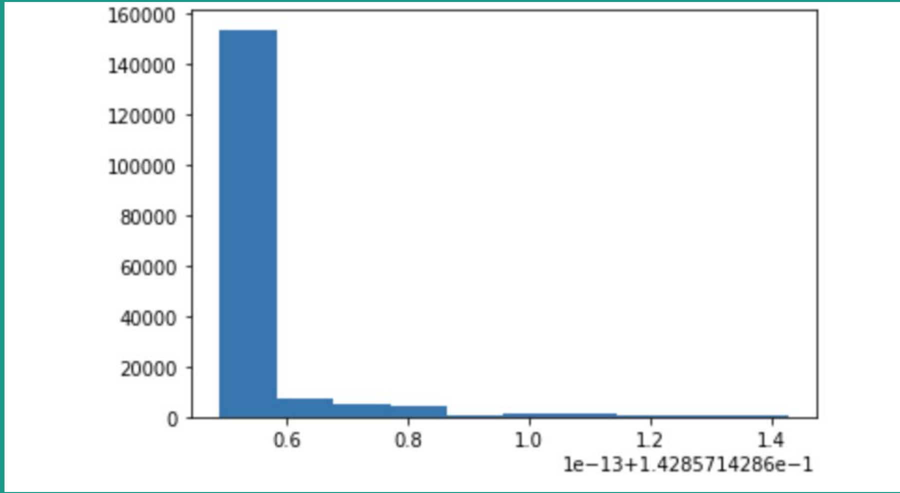
0	118682
1	9888
2	6882
3	5370
13	5180
4	4161
5	3527
12	3358
6	2993
11	2751
7	2690
8	2579
9	2451
10	2365
14	1610
15	176
16	99
24	67
36	64
20	52
17	33
52	32

(n=175 000)

Opened number of “Zeitbrief” in three months

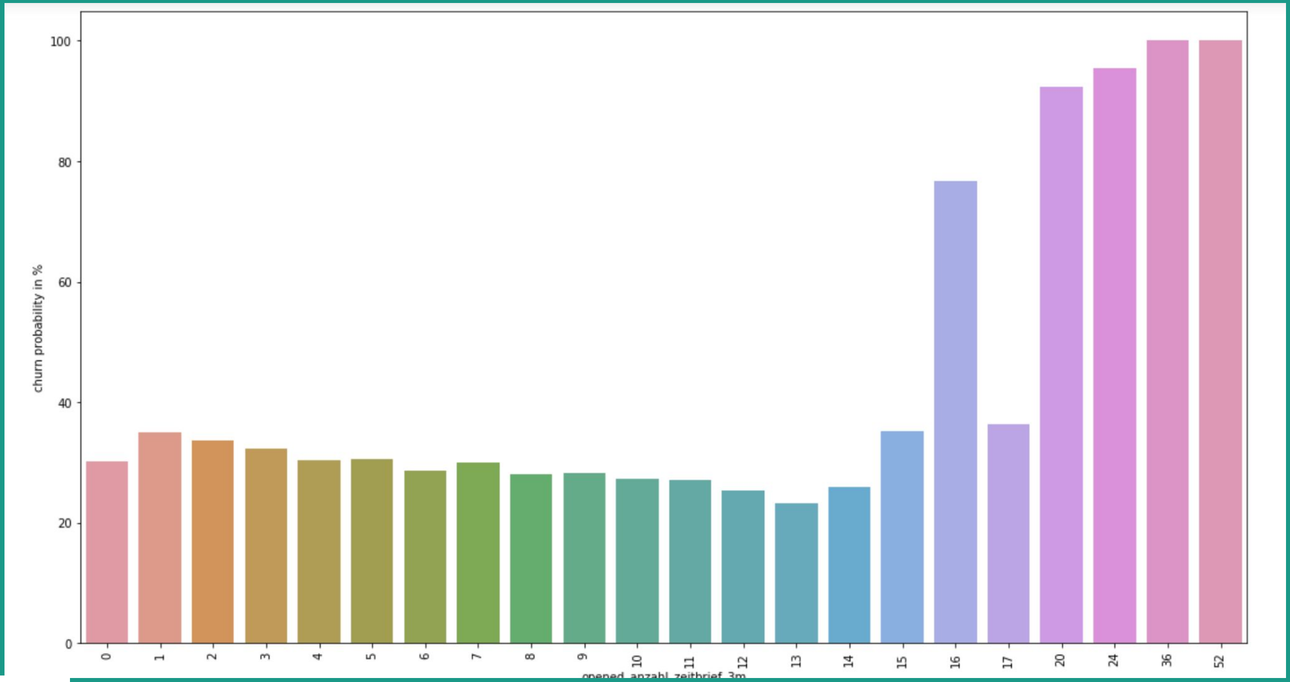


Logarithm of opened_anzahl_zeitbrief_3m



Boxcox of opened_anzahl_zeitbrief_3m with parameter -7

Opened number of “Zeitbrief” in three months



Churn probability by opened_anzahl_zeitbrief_3m: **Seems less relevant**

Numerical tests to answer these questions

Small numerical tests with a subset of the features yielded the following results:

Results of small numerical tests

- plz_3 as 697 dummies instead of no plz improves ROC_AUC by
  1-1.5 percentage points for logistic regression
  5 percentage points for K nearest neighbors

- plz_3 vs. plz_1 or plz_2 improves performance by up to
  1 percentage point (logistic regression)
  3.2 percentage points (K nearest neighbors)

- abo_registrierung_min_year and nl_registrierung_min_year binned and turned into dummies gives slightly better results than as naive numerical variable

- Treatment (log, scaling, drop) of extremely skewed features has little impact

Model building

Erster Punkt

Text hier einfügen Text hier einfügen Text hier einfügen Text hier einfügen Text hier einfügen Text hier einfügen

Text hier einfügen

Text hier einfügen Text hier einfügen Text hier einfügen Text hier einfügen Text hier einfügen Text hier einfügen.

Text hier einfügen Text hier einfügen Text hier einfügen Text hier einfügen Text hier einfügen Text hier einfügen

Zweiter Punkt

Abschließender Punkt

Beschreibung desselben in einer Zeile

"Dies ist ein sehr bedeutendes Zitat."

– Ein Experte

Dies ist der Ort für die Hauptaussage, die jeder aus dieser Präsentation für sich mitnehmen sollte.