Die ZEIT: churn or no churn

Capstone Project Presentation - neuefische Data Science Bootcamp

Team: Carlotta Ulm, Silas Mederer, Jonas Bechthold

Date: 26 November 2020

AGENDA

01 Introduction
Churn prediction for newspaper

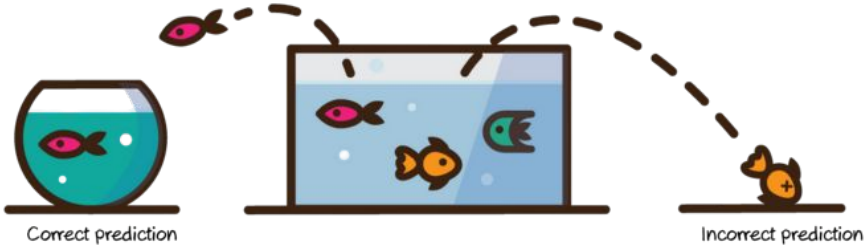
02 Data Insights
Dataset description
Who is a typical churn customer?

03 Machine Learning
Baseline model, feature selection & model improvement

04 Artificial Neural Network
Deep Neural Network
Training and Tuning

05 Model Deployment Recommendations

06 Future Work



Target

1. Predict subscribers that will churn in near future.

2. Show interesting data insights and give recommendations for churn prevention (e.g. time)

Metrics

Approach: Binary classification

Metrics (churn):

- **Recall:** Identify as many as possible real subscription churns

conflict

- **Precision:** Avoid too many disturbing mails to loyal subscribers

Image Source

01 Introduction

Publishing

Print and online

Customer Relation

Revenues, planability, stability

Churn

Early detection and prevention

01 Introduction - Dataset

Data

- Size ≈ 210.000
- 177 features
- Subscription orders from >2012
- Churn period 2019 - 2020

Restriction

- max 4 subscriptions per Household
- Size ≈ 184.000
- Loss of 12 %

Feature Overview

Customer Information

- title
- city/metropolitan
- country
- zip codes

Time Information

- reading time
- churn date
- newsletter

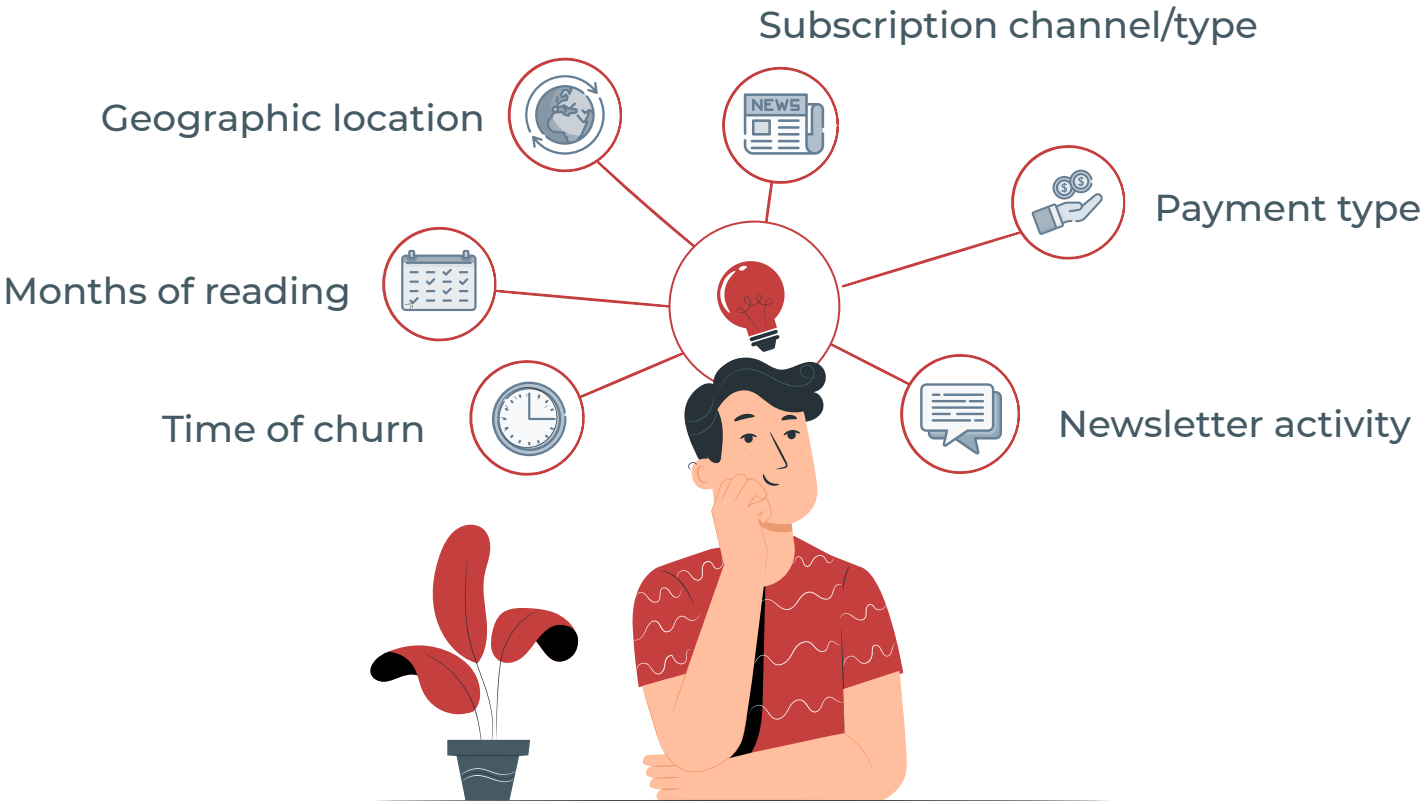
Subscription Information

- channel
- subscription type
- payment
- billing/student...

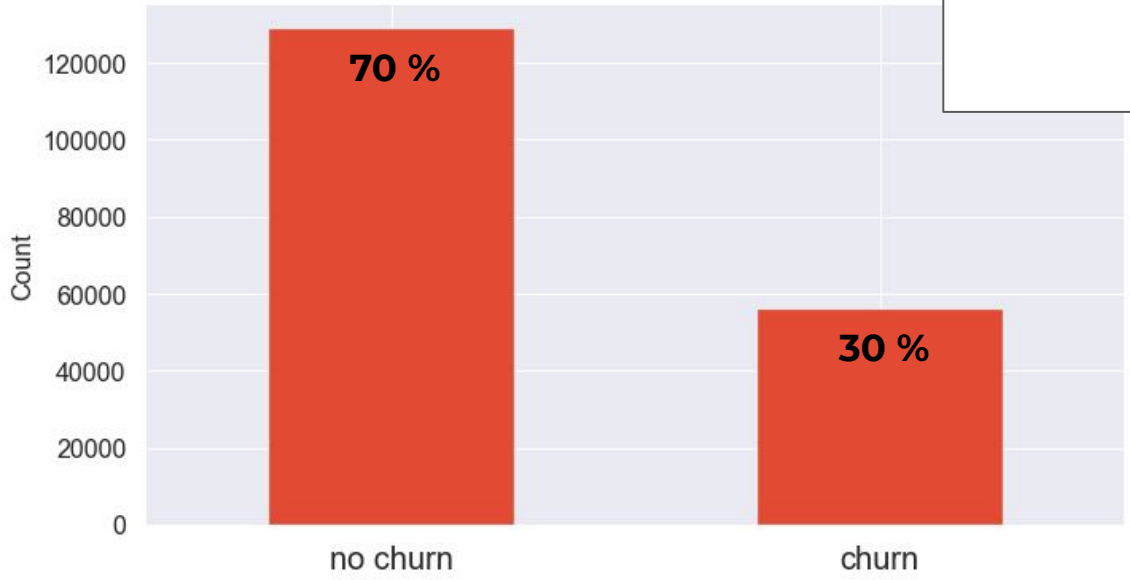
Activity Information

- newsletter
- shop buys
- opens/clicks etc.

02 Data Insights



02 Data Insights - Churn rate

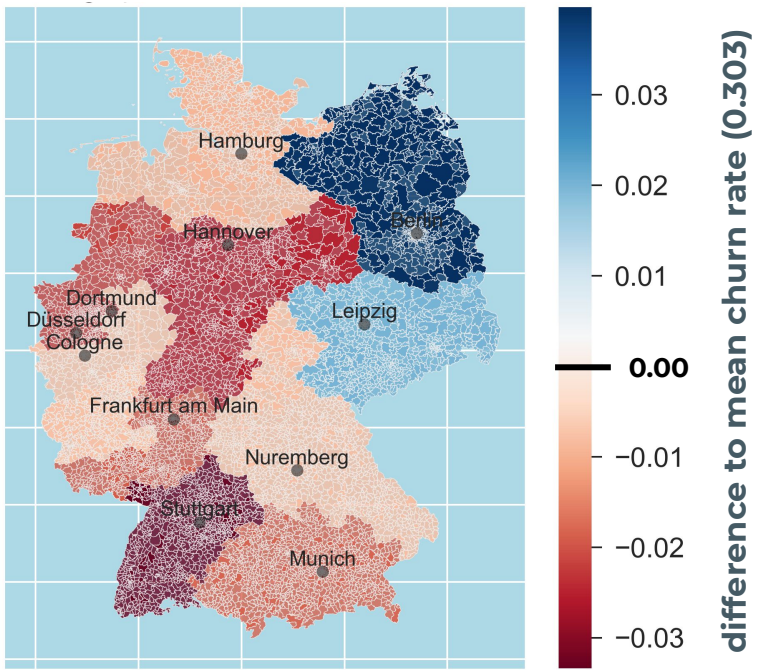


"If I always choose churn, I have an **accuracy** round 0.303 and a **F1 score** of 0.352."

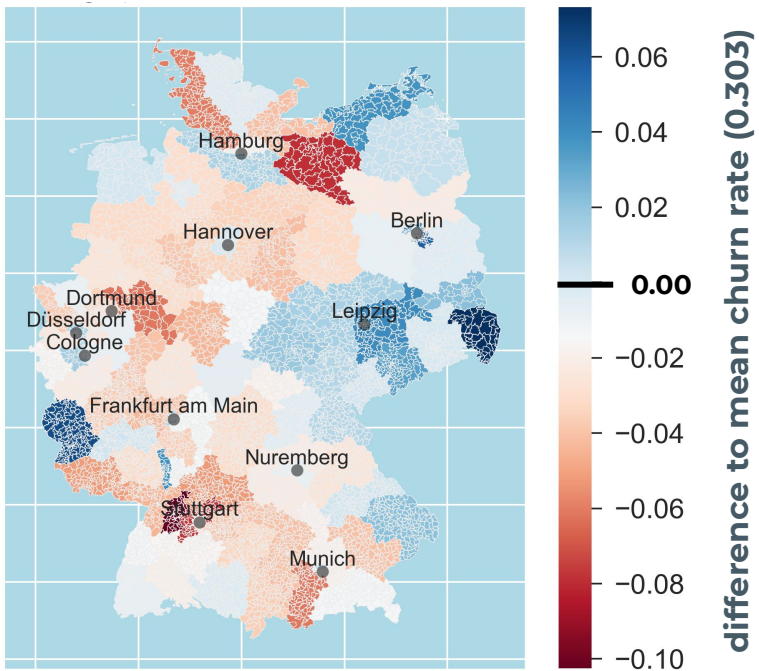
- Naive Predictor 2020

02 Data insights - Geographical Representation

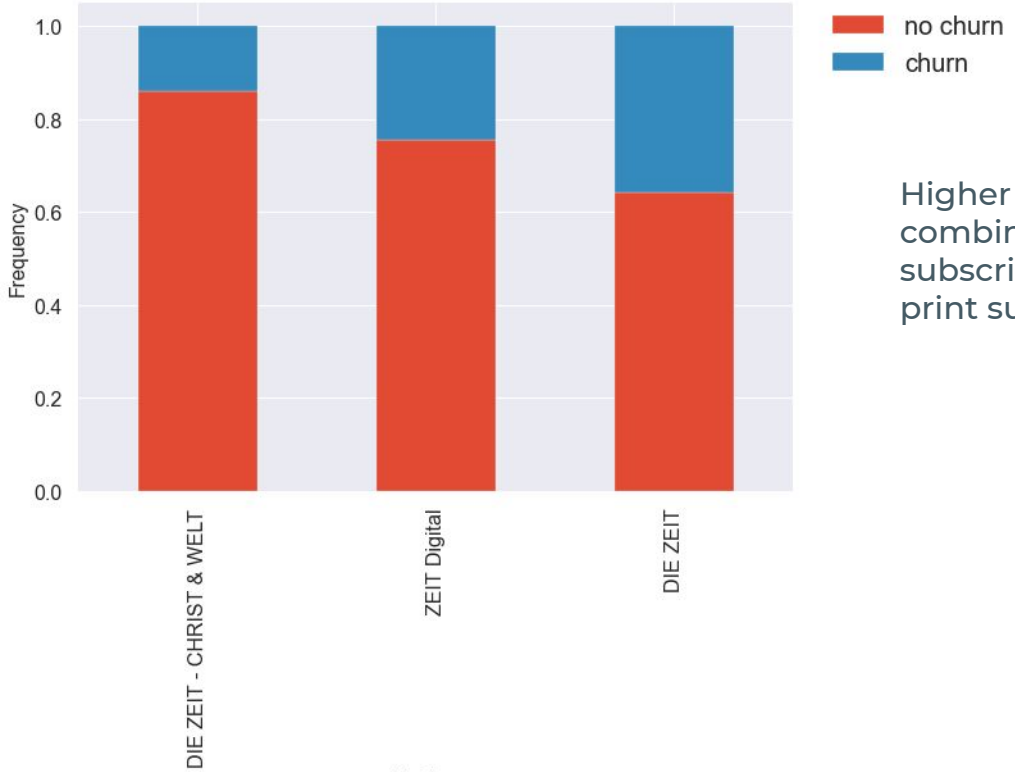
1 digit zip code (e.g. 2xxxx)



2 digit zip code (e.g. 21xxx)

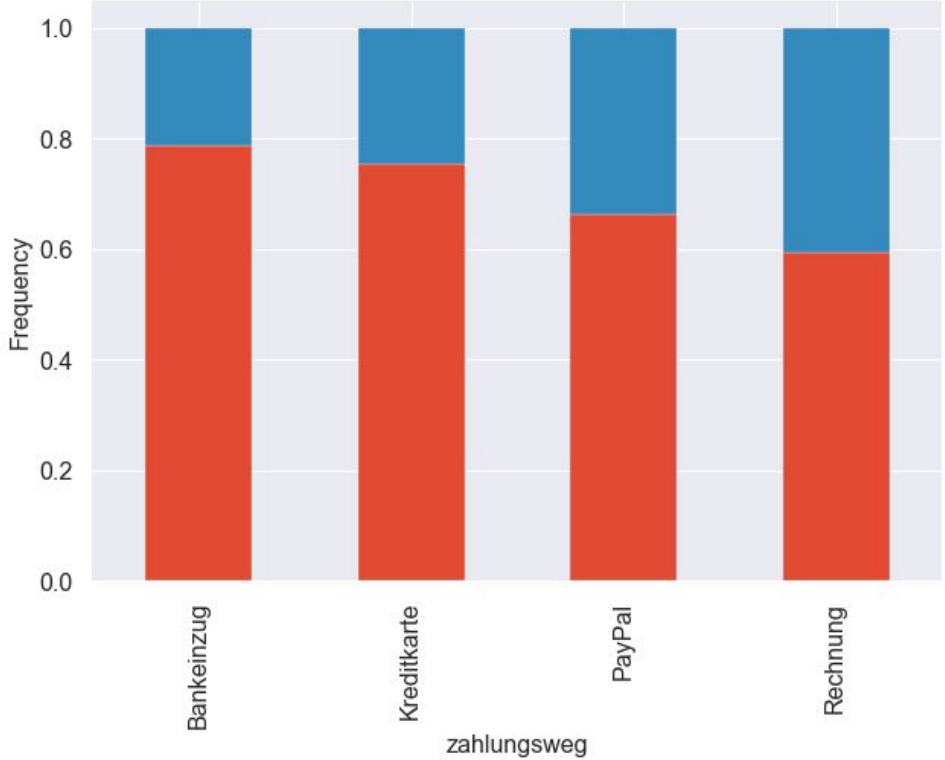


02 Data insights- Subscription type



Higher percentage of loyal customers for combined print & Christ and Welt subscription than for the digital or only print subscription.

02 Data insights - Payment

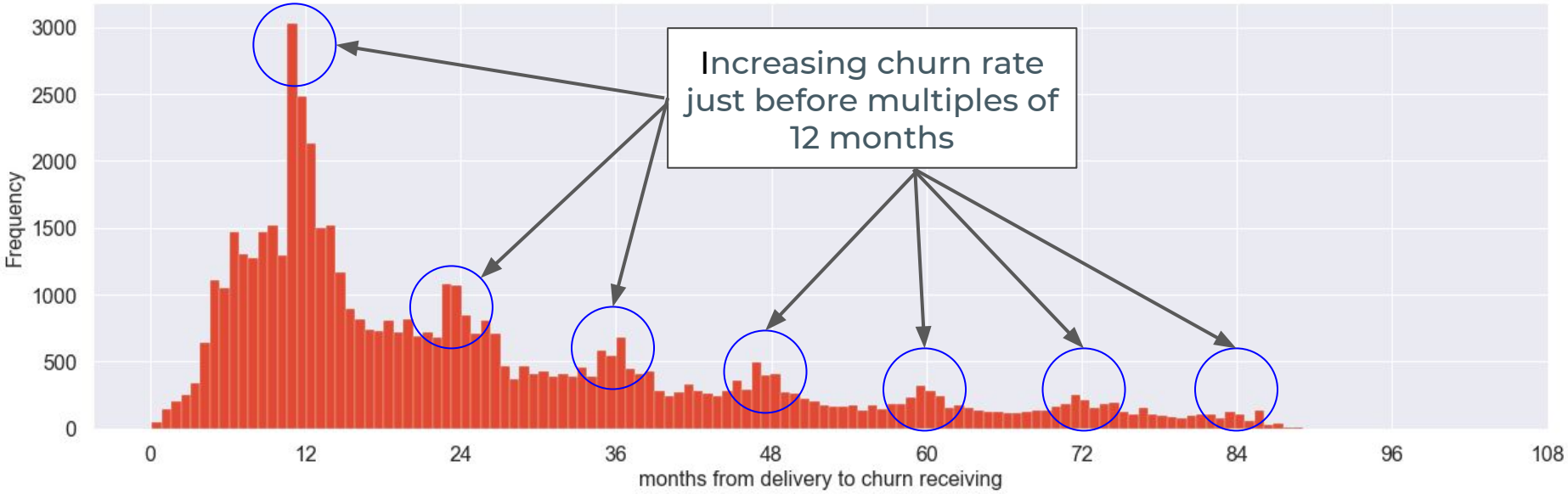


Payment type has an impact on the churn rate:

- Bankeinzug: people are financially secured
  → 22 % churn

- Rechnung: people who want to have control over their payments
  → 41 % churn

02 Data insights - Time of churn



02 Whole Dataset -- Churners

Gender

43.3% 56.7%

Gender

47% 53%

Subscription

54.8% 41.7% 3.5%

Subscription

64.8% 33.6% 1.6%

Payment method

Direct debit : 52 %
Invoice: 45.3%
Credit card: 1.7%
Paypal: 1%

Payment method

Direct debit : 58.4 %
Invoice: 39%
Credit card: 1.4%
Paypal: 1,2%

02 Whole Dataset -- Churners

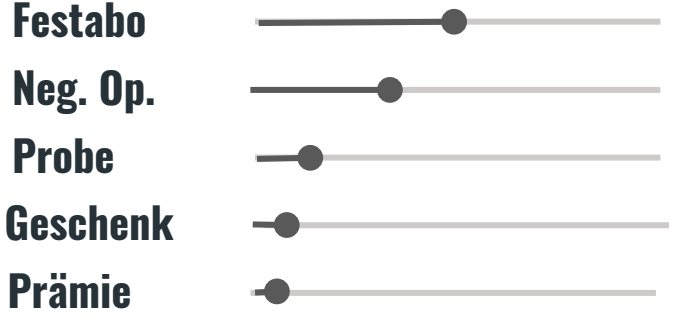
Studentenabo

11%

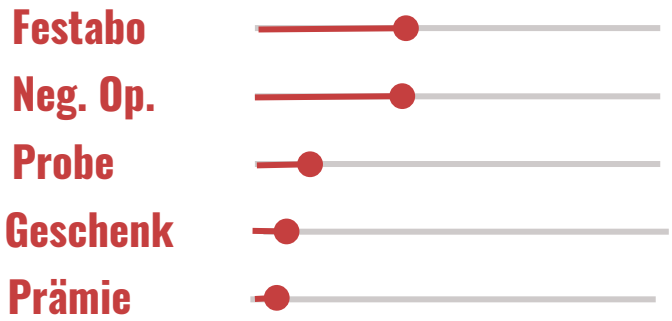
Studentenabo

16%

Subscription type



Subscription type



03 - Machine Learning

Baseline

- Preprocessing
- Different ml methods
  - LogReg
  - KNN
  - SVC
  - RF
  - AdaBoost
  - XGBoost

Optimization

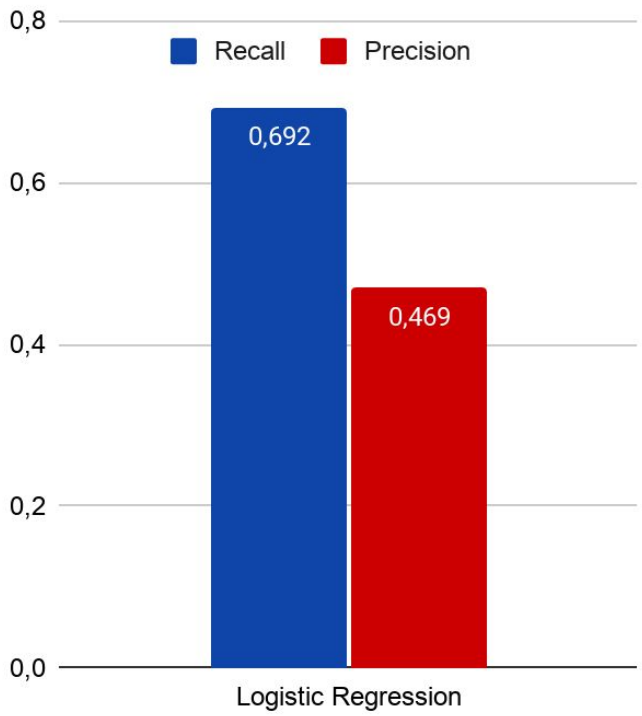
- Feature engineering
- Feature selection
- Random search
- Grid search
- Model combination

Reflection

- Evaluation
- Post processing

Baseline model!

Raw data different approaches



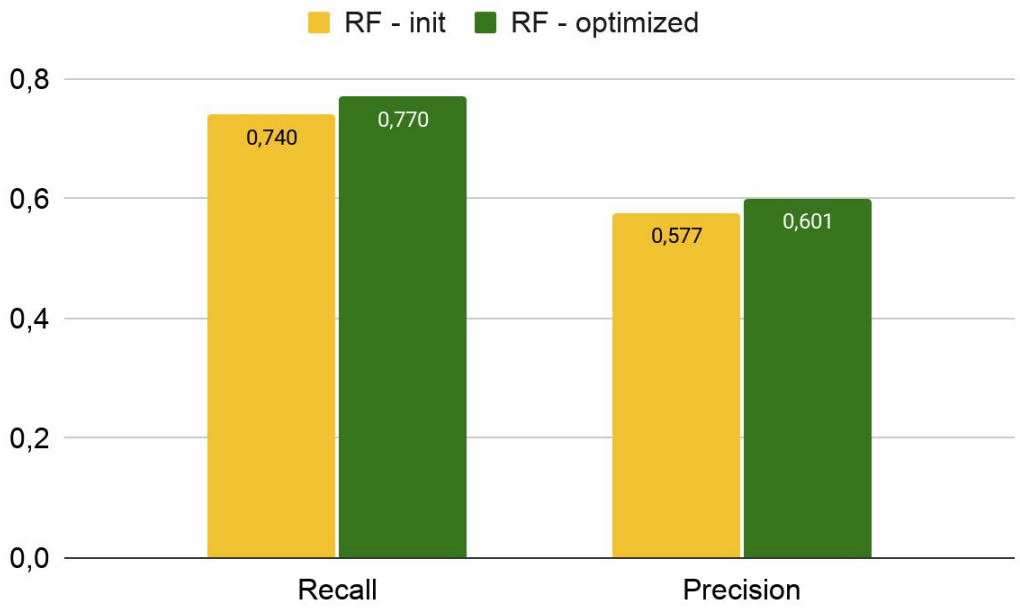
n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Recall: TP/actual yes = 100/105 = 0.95
Precision: TP/predicted yes = 100/110 = 0.91
F1: weighted average of recall and precision

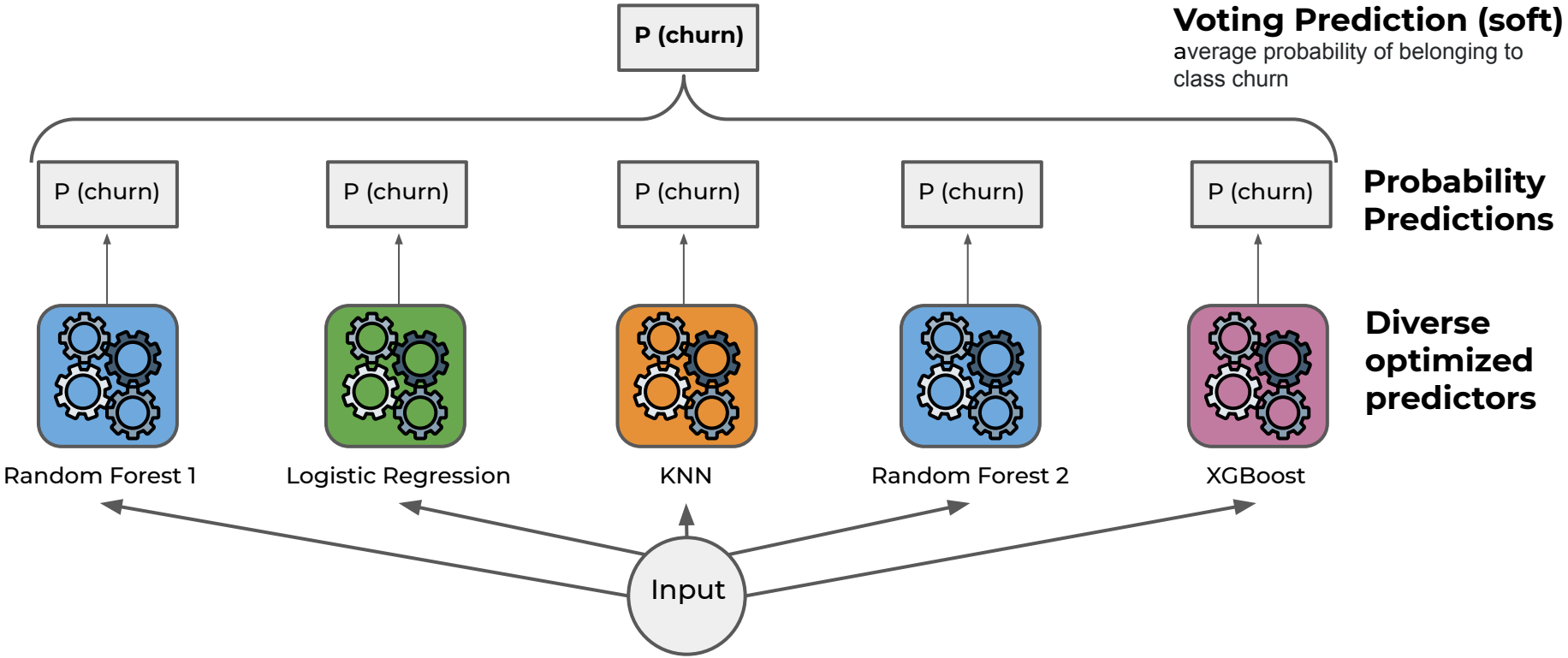
03 - Best Single Classifier - Random Forest

Optimization:

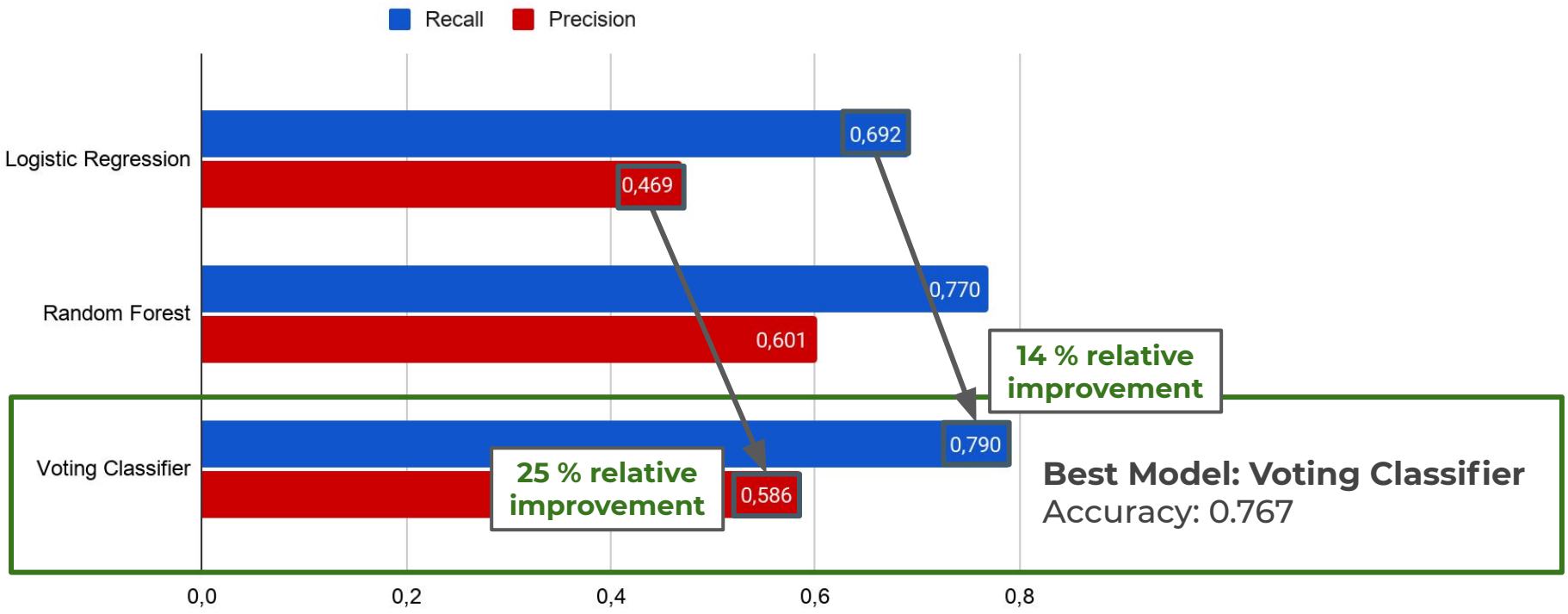
1. **Feature Selection**
2. **Majority downsampling** (no churn) for data balance
3. **Randomize Search** to explore and evaluate hyper parameter space (wide grid)
4. **Grid Search** to search for optimal parameters (narrow grid)



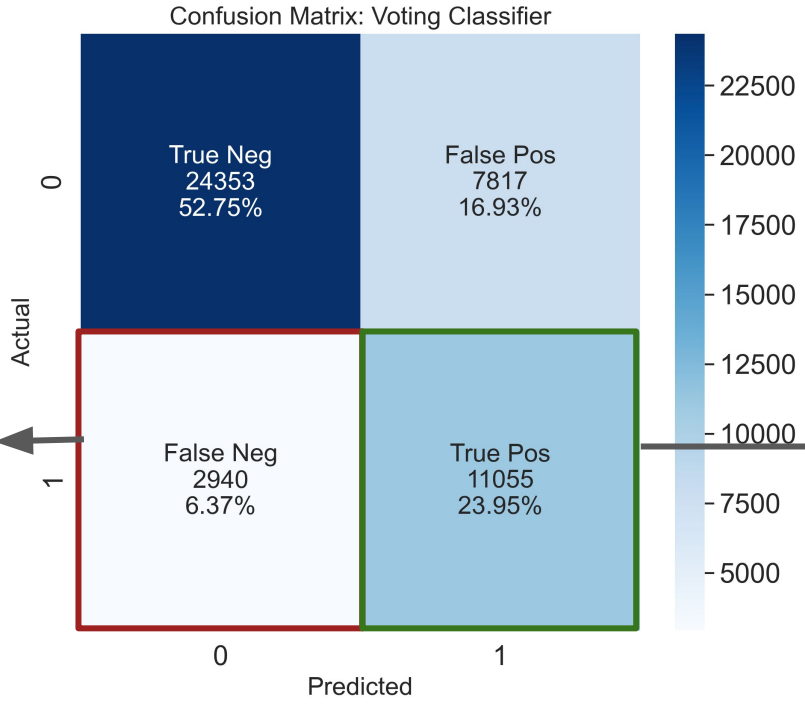
03 - Best Classifier - Voting



03 - Best Classifier - Comparison



03 - Voting Classifier - Results



Possible additional revenue

Possible Revenue by churn prevention

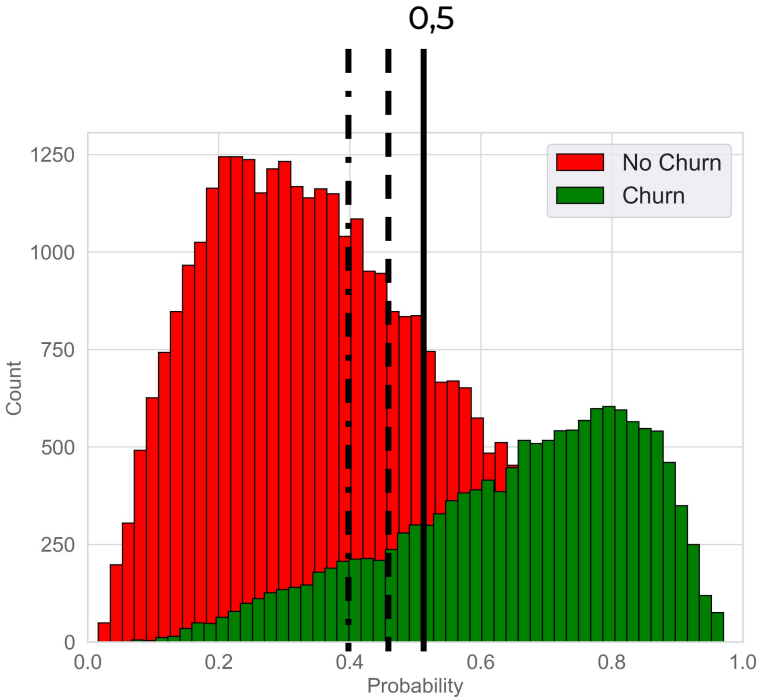
Correct predictions 76,7%

False predictions 23,3%

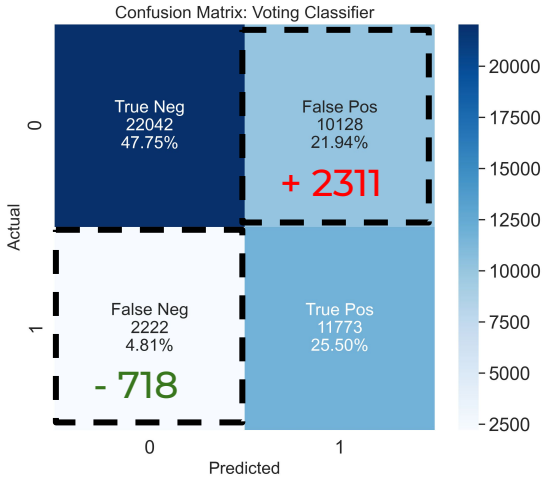
3.1 Million €

- If all churns can be prevented, (price 280 € p. Sub.)
- Only 30 % test set considered here

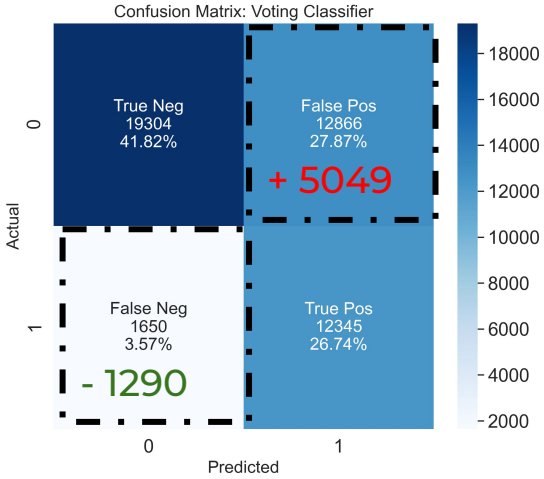
Voting Classifier - Probability Threshold



P (churn) ≥ 0.45



P (churn) ≥ 0.40



03 Voting Classifier -Analysis of FN

Gender

41.5%

58.5%

Studentenabo

6.3%

Subscription

53.7%

43%

3.3%

Subscription type

Festabo

Neg. Op.

Probe

Geschenk

Prämie

Payment method

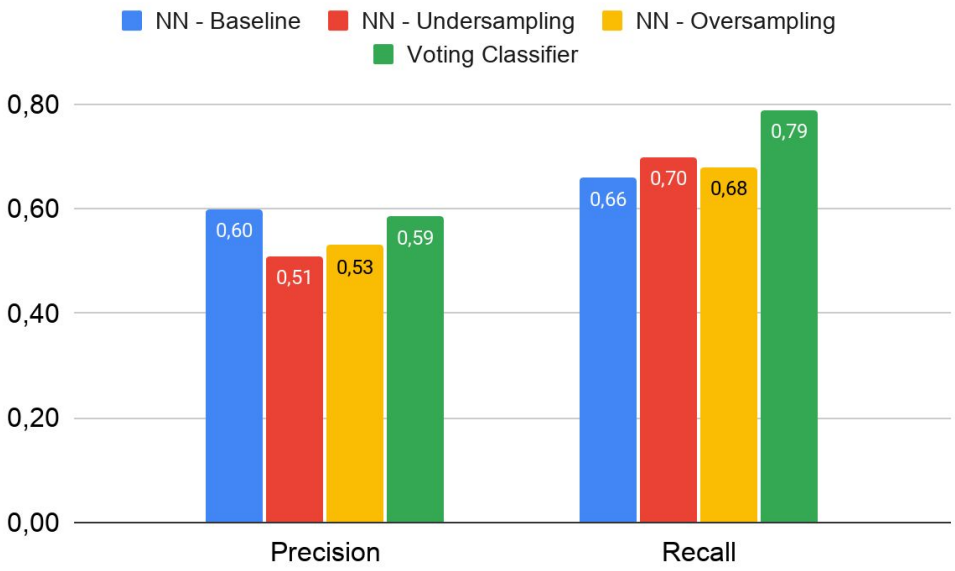
Direct debit : 63.5 %

Invoice: 33%

Credit card: 2.6%

Paypal: 0.9%

04 Artificial Neural Network



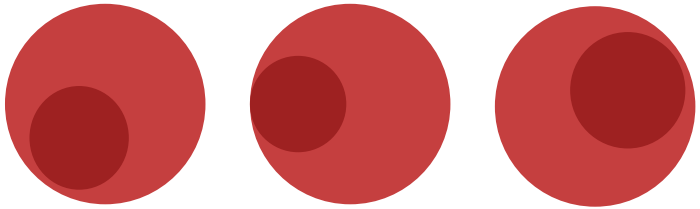
Correct predictions

70%
(76,7%)

False predictions

30%
(23,3%)

Churns Predicted = Recall

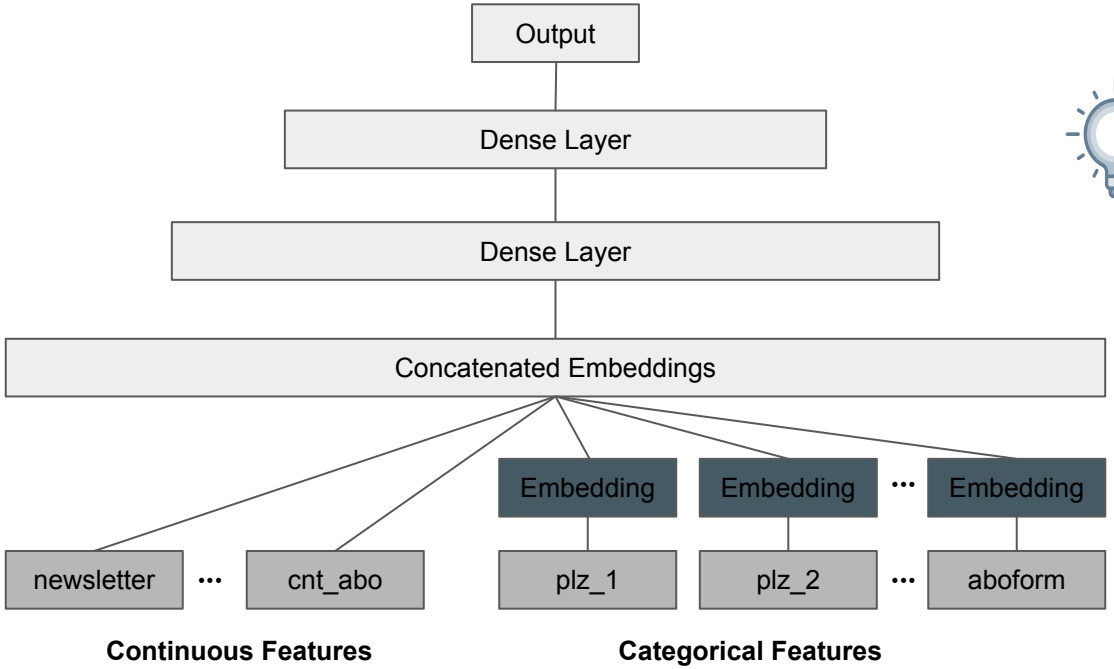


66%
Baseline

69%
Oversample

71%
Undersample
(79%)

04 Artificial Neural Network - Embedding



Correct predictions	False predictions
72,1% (76,7%)	27,9% (23,3%)

Churns Predicted

70,2%

(79%)

05 Deployment

plan

Load: dataset

We use a csv file

do

Start: churn-predictor.py

The program will ask for inputs

check

Take: Predictions

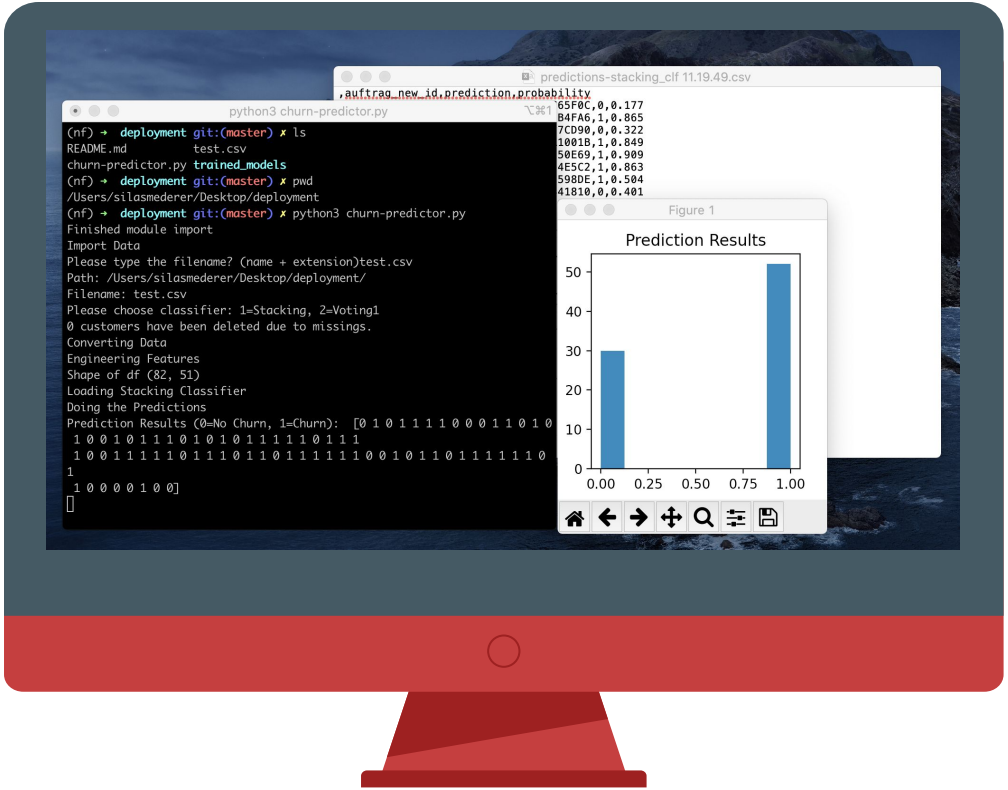
Datafile is exported to file directory

act

Contact: Customers

Motivate detected customers to stay

05 Deployment

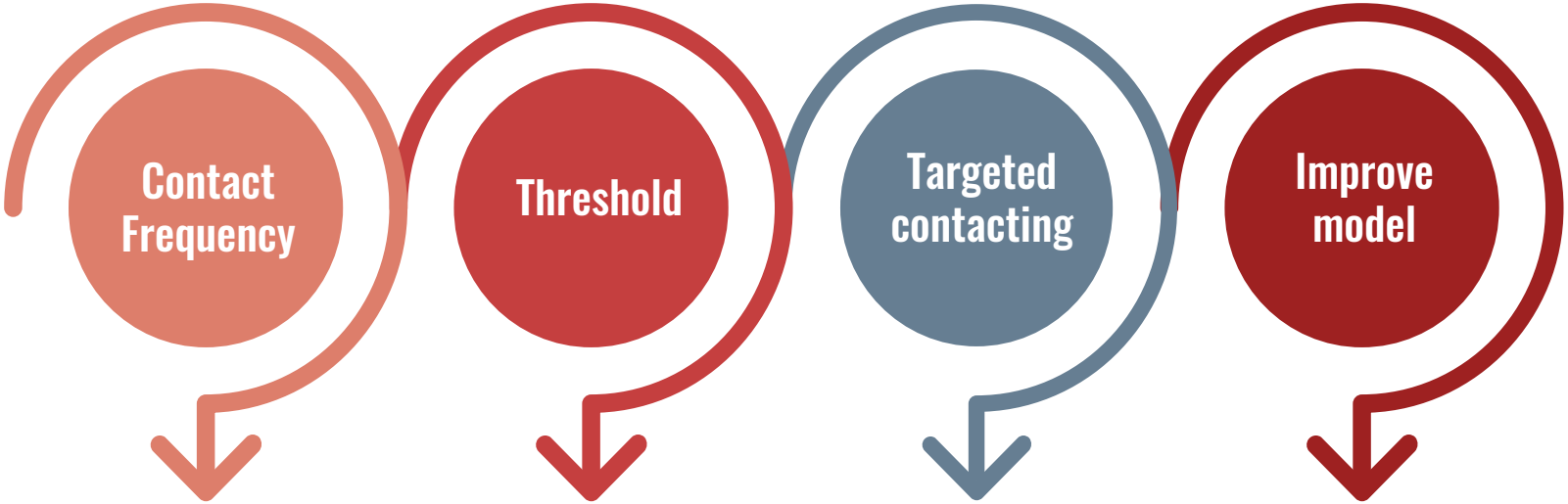


Input:
dataset

Output:
predictions
& proba-plot

Options:
Stacking
Voting

05 Recommendations



After 9 months from Subscription start

Change probability and prediction focus

Create incentives for different user groups

Use time related features to build a better model

06 Future Work

More & different data

- Time relation
- Online activities

Machine Learning

- Tune ANN or try CNN
- Error Analysis

Subscriber specific churn prevention

- Develop incentives to keep customers
- Personalize mail contact

Image Source

THANKS

Silas

Academic background in nature science looking for jobs in solutions engineering or consulting
LinkedIn: silas-mederer
GitHub: sls-mdr

Carlotta

Graduate economist with a passion for data science and machine learning.
LinkedIn: carlotta-von-ulm-erbach
GitHub: carlotta-ulm

Jonas

Data scientist with research experience and engineering background.
LinkedIn: jonas-bechthold
GitHub: jb-ds2020