# Project Write-up – Predicting Conversions from Print Newspaper Subscription Sales

## Abstract

The goal of this project was to train a classification model which predicts print newspaper subscription order conversions (=subscriptions reaching 90th day of regular paid subscription phase without being churned). In subscription businesses conversions are a central success metric for sales activities since subscriptions are regularly offered with a free trial phase upfront. A good performing model may help sales department to (1) identify subscription characteristics to focus on when selling subscriptions; (2) target subscription orders with subsequent "customer care actions" after purchase to increase conversion rates; (3) help sales and controlling gain a better idea on success of sales campaigns without delay after subscriptions are sold. My analysis was based on 3 months of print subscription sales between 01/07/21 - 30/09/21 of German daily newspaper publisher Frankfurter Allgemeine Zeitung. In feature engineering, I leveraged customer and subscription characteristics to create a set of mainly categorical and some numerical features. In terms of modeling, I used logistic regression, random forest, KNN, Gradient Boosting and naïve bayes (Bernoulli).

## Design

The project topic addresses one central question of (data-driven) sales activities. Feature data was collected from an internal source comprising data on the order, the purchased product, the customer itself and the customer's historical relationship to the publisher. Target data was provided from within the company. Extracting insights on order properties (such as payment periods, marketing channels and customer age) contributing to high conversion probabilities via machine learning models may support sales staff to sell more subscriptions. It may help to focus sales activities to reduce early churn of subscriptions and thus contribute to overall profit. Precision was chosen as central evaluation metric since additional outreach to customers incurs substantial time/service costs.

## Data

The dataset contains of 13,015 orders (approx. 15% of which are conversions) with 31 features for each, 26 of which are categorical. A few feature highlights include "marketing channel", "payment period", "payment method", "paid trial / free trial", "number of orders in last year" or "days since first order in customer relationship". Nearly 70% of individual features were grouped into broader categories and investigated by creating dummy variables. The final feature set contained 81 variables.

## Algorithms

*Feature Engineering*

1. Converting subscription data into usable format with one line per order
2. Creating bins of broader categories for individual feature entries (such as "rebate groups" or "sub types")
3. Generating target variable for each subscription based on duration and type of subscription
4. Generating features on historical customer relationships (such as time since first order)
5. Converting categorical features to binary dummy variables (such as age group)
6. Check for multicollinearity using VIF and deleting highly correlated features

*Models*

Logistic regression, KNN, Random Forest, Gradient Boosting, Naïve Bayes (Bernoulli), Ensemble Stacking and Ensemble Voting were used before settling on Gradient Boosting as the model with strongest precision performance. Throughout, feature coefficients (logistic regression), feature gain, permutation importance and SHAP values were used to check for significance of each feature.

*Model Evaluation and Selection*

The entire training dataset of 13,015 records was split into 80/20 train vs. test, and best model was selected using a 5-fold stratified cross validation on the training portion only. Due to class imbalance (15% vs. 85%) oversampling and class weights were applied to improve performance. Predictions on 20% test were limited to the very end, so this split was only used and scores seen just once. Models were evaluated based on average precision. The best performing model (Gradient boosting) had a precision of 0.81 on the test sample versus a mean precision of 0.86 on the 5-fold CV sample.

## Tools

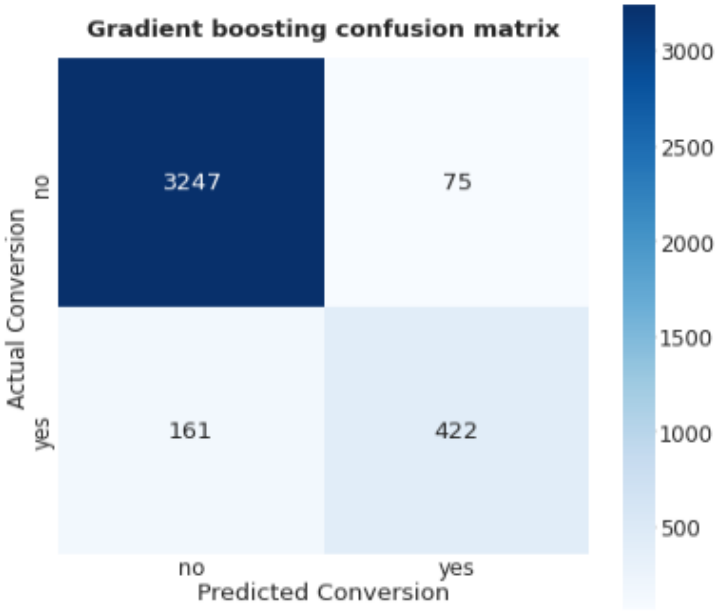
- NumPy and pandas for data manipulation
- Scikit-learn, xgboost and mlxtend for modeling
- SHAP and statsmodels for feature evaluation
- Matplotlib and seaborn for plotting
- Tableau for visualization (link)

## Communication

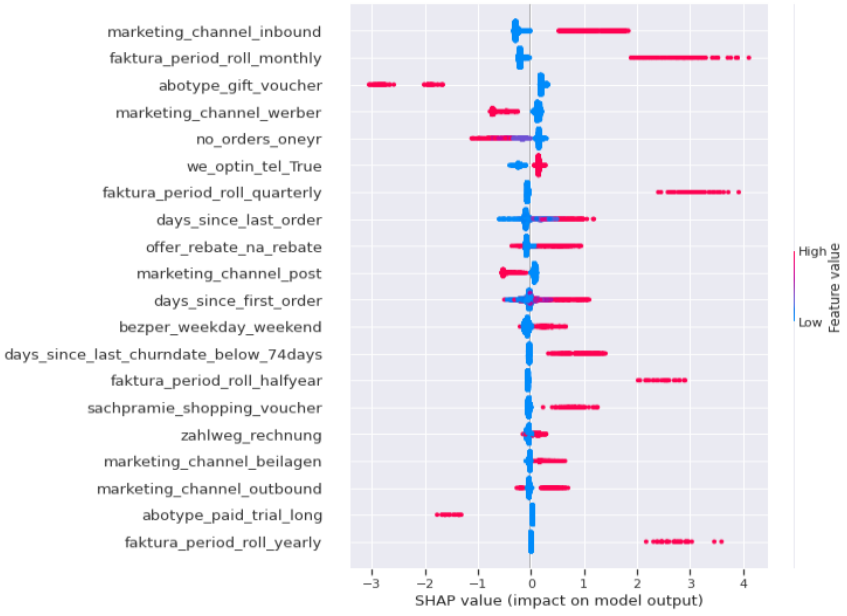
In addition to the slides and visuals presented, project outputs will be embedded on my personal github.

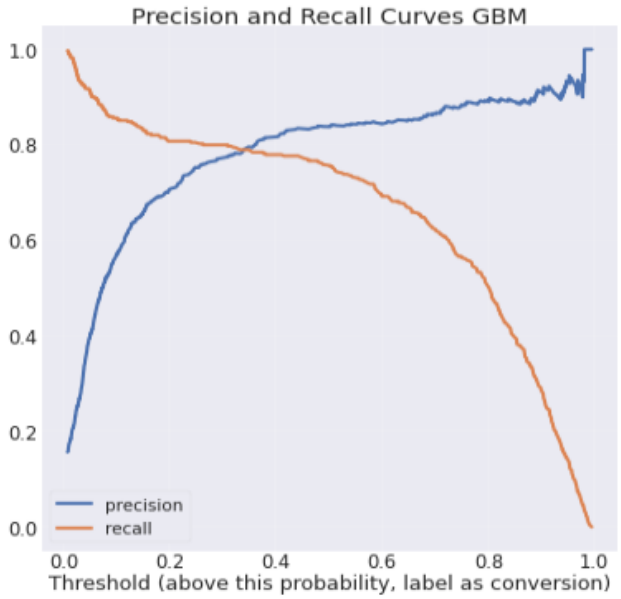
| | features | vif |
|---|---|---|
| 75 | days_since_last_enddate_below_102days | 4.869430 |
| 51 | marketing_channel_inbound | 4.572259 |
| 73 | offer_rebate_na_rebate | 3.855020 |
| 40 | we_optin_tel_True | 3.783352 |
| 48 | marketing_channel_coop | 3.631829 |
| ... | ... | ... |
| 49 | marketing_channel_dig_na | 1.018086 |
| 31 | sachpramie_household | 1.015375 |
| 25 | zahlweg_credit_card | 1.014030 |
| 26 | zahlweg_other | 1.010953 |
| 36 | sachpramie_travel | 1.008196 |

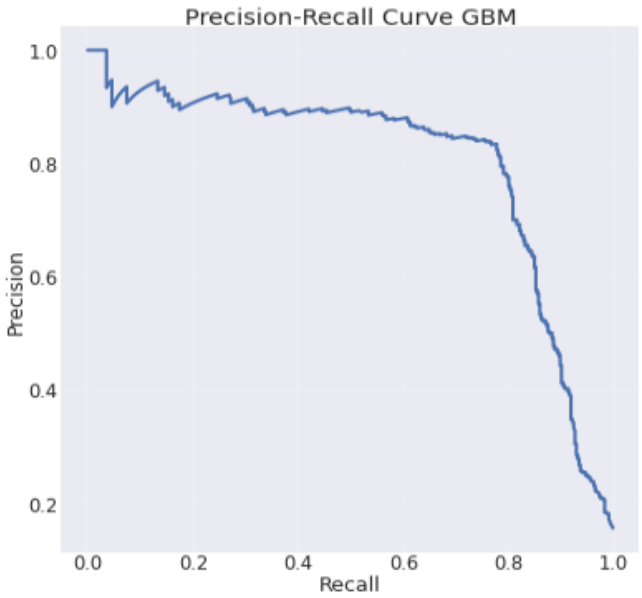
81 rows × 2 columns

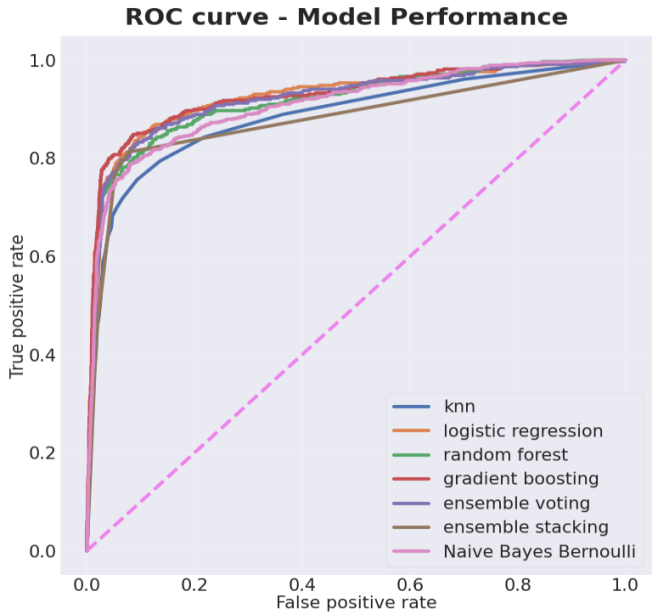



| | feature | value_gain |
|---|---|---|
| 0 | marketing_channel_inbound | 100.457344 |
| 1 | abotype_gift_voucher | 41.003235 |
| 2 | faktura_period_roll_monthly | 38.460331 |
| 3 | faktura_period_roll_quarterly | 28.940588 |
| 4 | days_since_last_churndate_below_74days | 28.899101 |
| 5 | faktura_period_roll_yearly | 19.551149 |
| 6 | faktura_period_roll_halfyear | 14.801983 |
| 7 | marketing_channel_werber | 12.701650 |
| 8 | sachpramie_gas_voucher | 9.350249 |
| 9 | offer_rebate_na_rebate | 8.665698 |
| 10 | abotype_student | 7.822011 |
| 11 | sachpramie_shopping_voucher | 7.720592 |
| 12 | marketing_channel_onsite | 7.644912 |
| 13 | active_digsubs_atorder | 7.279757 |
| 14 | no_orders_oneyr | 7.165772 |






```
Precision Scores:
Gradient Boosting =  0.8385269121813032
Logistic Regression =  0.6509240246406571
Ensemble Voting =  0.7730870712401056
Random Forest =  0.6810933940774487
Naive Bayes Bernoulli =  0.738831615120275
KNN =  0.8104265402843602
Ensemble Stacked =  0.7228915662650602
```