



DIE  ZEIT

Churn Prediction

Beat the logistic regression model of 'DIE ZEIT'

Authors: Sebastian Fuhrer, Marco Zagermann

WHAT IS CHURN PREDICTION?

WHAT IS CHURN PREDICTION?



PROBLEM

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

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)



WHAT IS CHURN PREDICTION?

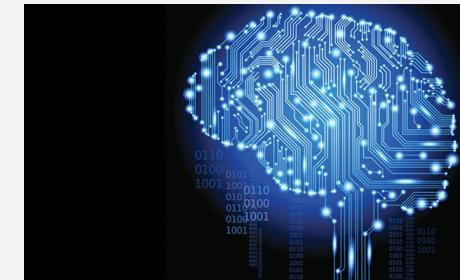


PROBLEM

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.)

WHAT IS CHURN PREDICTION?

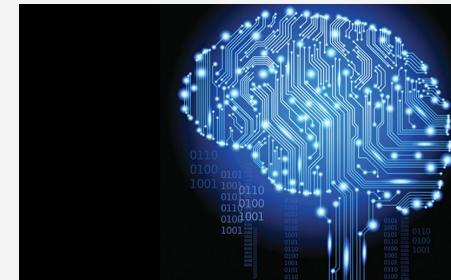


PROBLEM

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

02
EXPLORATORY
DATA ANALYSIS

03
FEATURE
SELECTION

04
MACHINE LEARNING
MODELS

05
FUTURE WORK

01

THE DATA SET



THE DATA SET

BASIS

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

THE DATA SET

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

THE DATA SET

BASIS

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

DIMENSIONS

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

VARIABLES

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

THE DATA SET

BASIS

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

DIMENSIONS

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

VARIABLES

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

CHALLENGE

Pre-existing model:

- Logistic regression
- Lasso / Ridge feature selection

THE DATA SET

	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

DIMENSIONS

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

VARIABLES

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

CHALLENGE

Pre-existing model:

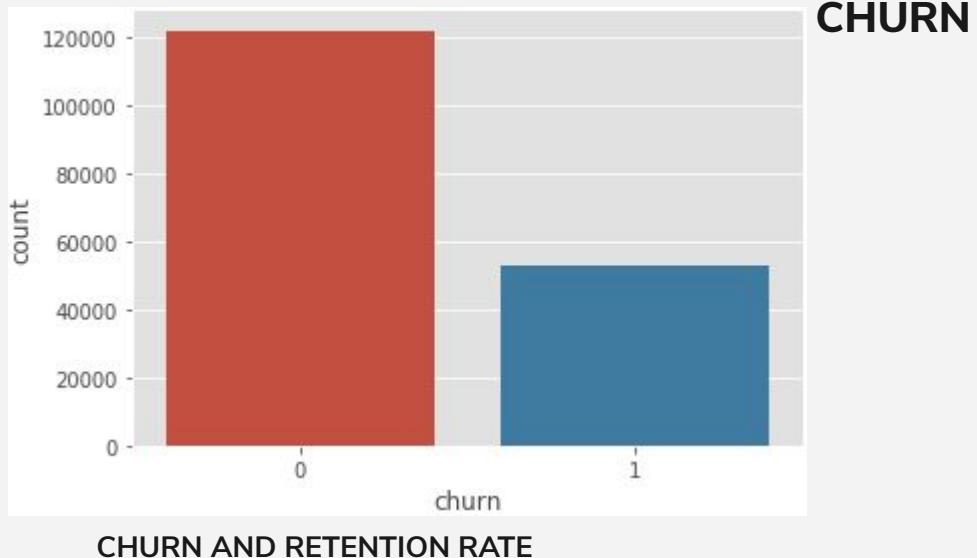
- Logistic regression
- Lasso / Ridge feature selection

A photograph showing two men in an office environment. One man, wearing a plaid shirt and jeans, stands on the left facing right, gesturing with his hands as if in conversation. The other man, wearing a white button-down shirt and dark pants, stands across from him, holding a white coffee cup. They are positioned in front of a large window that looks out onto a brick building. In the foreground, there's a wooden desk with various office supplies like pens and a small potted plant. The floor is made of light-colored wood.

02

EXPLORATORY DATA ANALYSIS

CUSTOMER BEHAVIOR

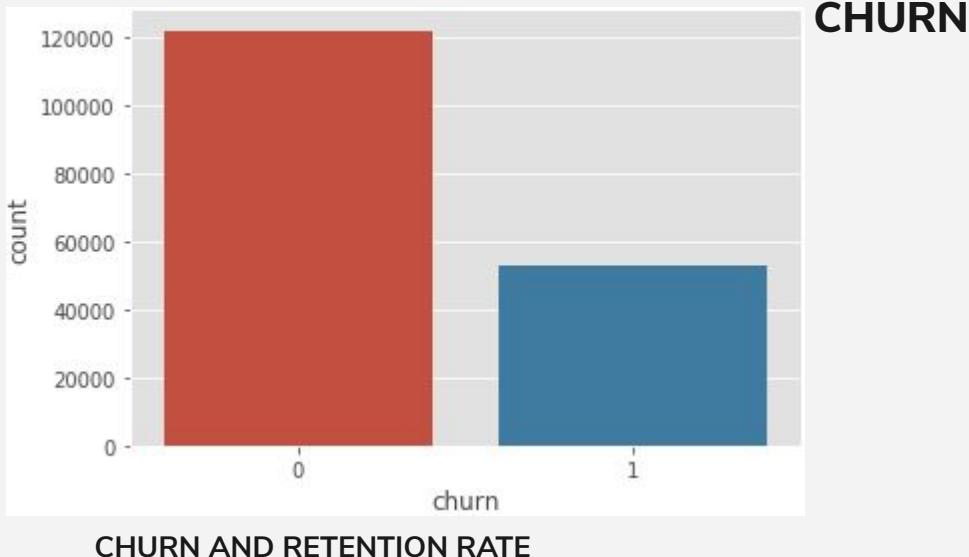


69.78%
RETENTION

30.22%
CHURN

CUSTOMER BEHAVIOR

- Reference period:
06/2019-05/2020



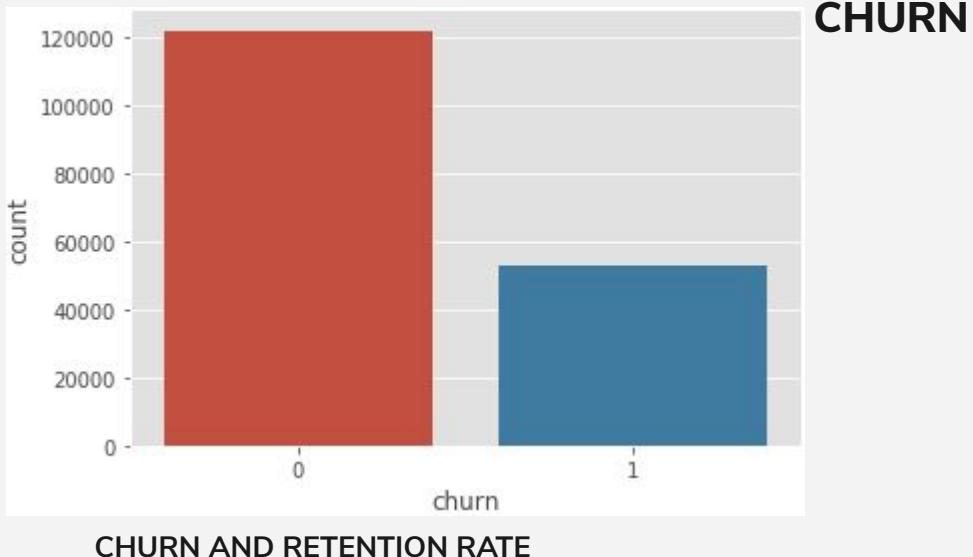
CHURN AND RETENTION RATE

69.78%
RETENTION

30.22%
CHURN

CUSTOMER BEHAVIOR

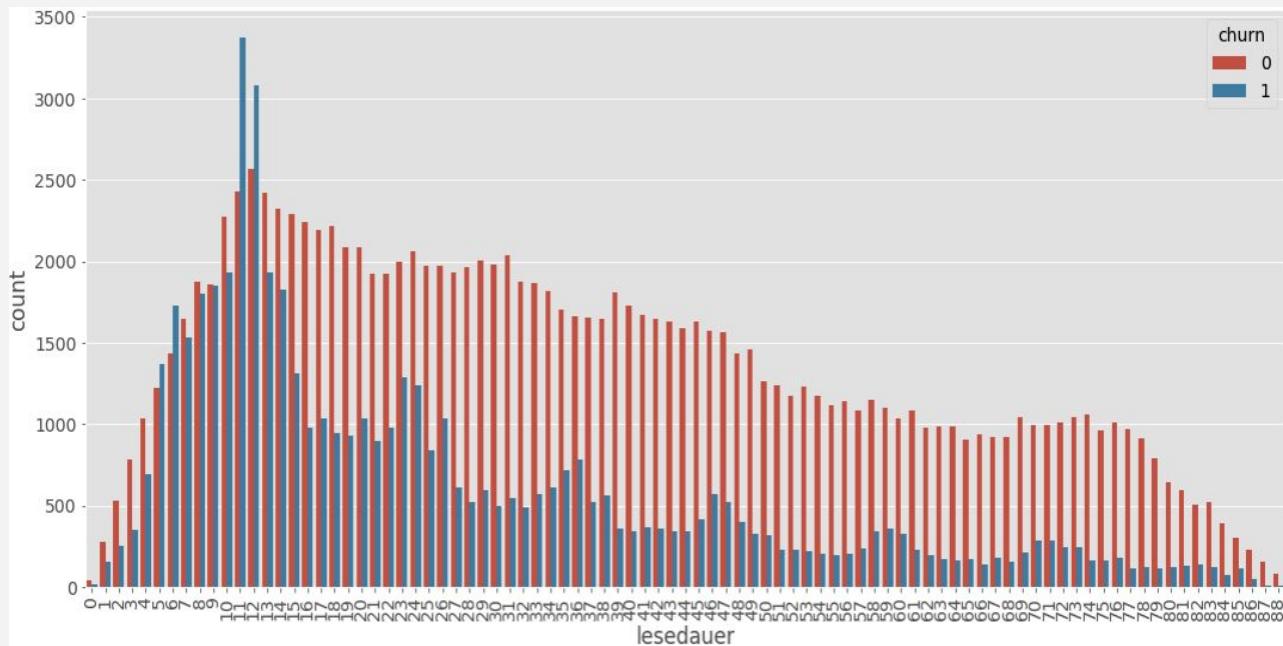
- Reference period:
06/2019-05/2020
- Unbalanced distribution of
target variable ‘churn’



CHURN AND RETENTION RATE

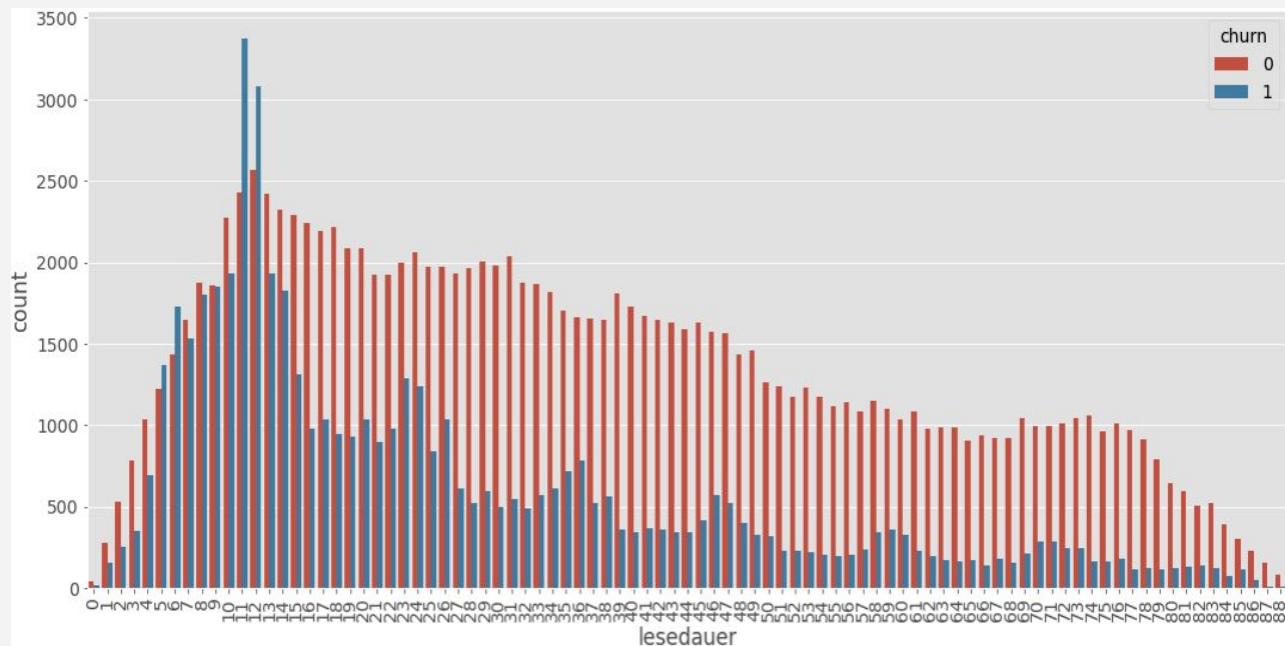


MONTHS OF READING



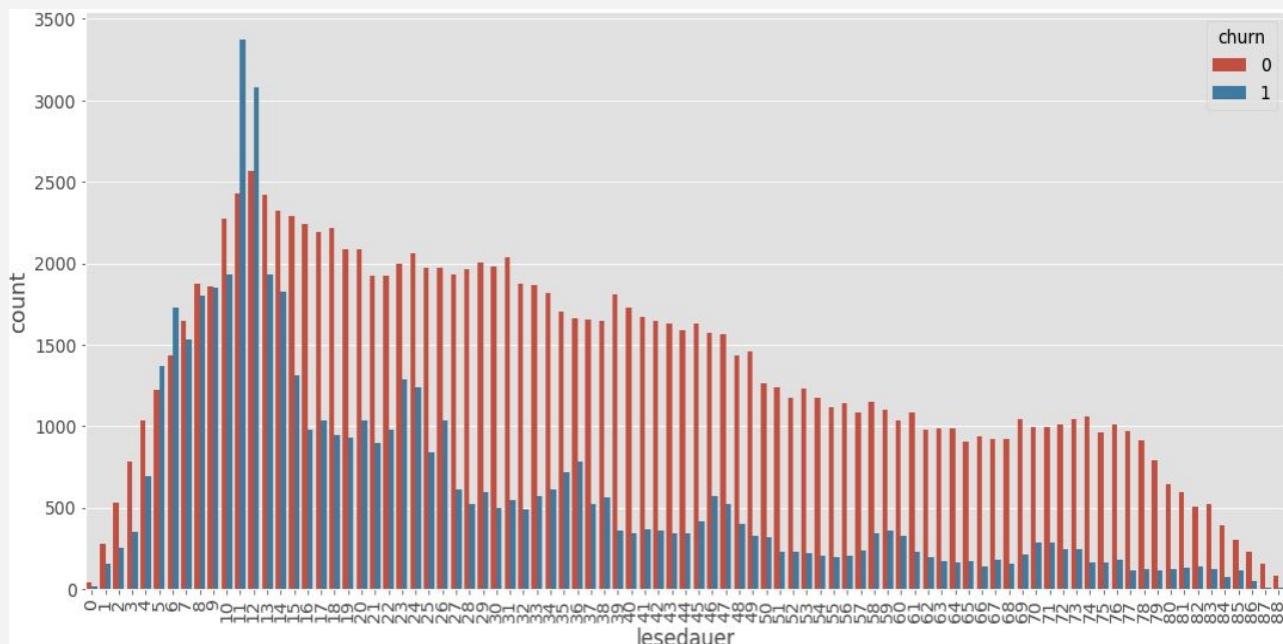
MONTHS OF READING

- Churn rate decreases with increasing reading time



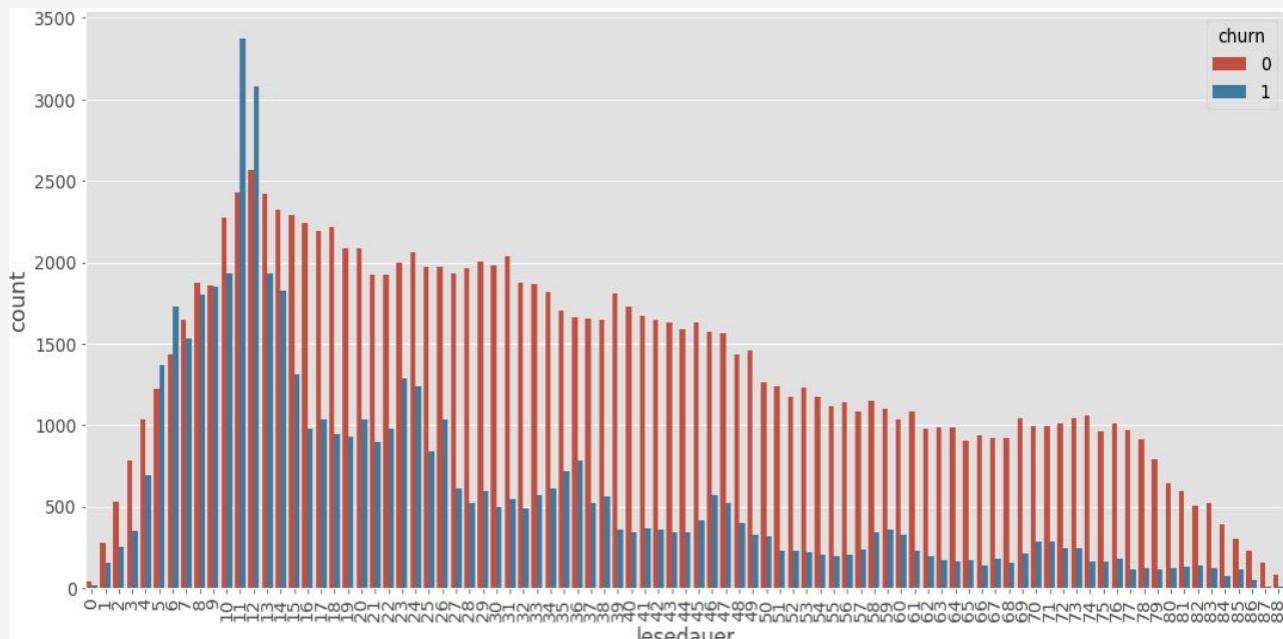
MONTHS OF READING

- Churn rate decreases with increasing reading time
- 12 months of reading is the major threshold towards loyal customers



MONTHS OF READING

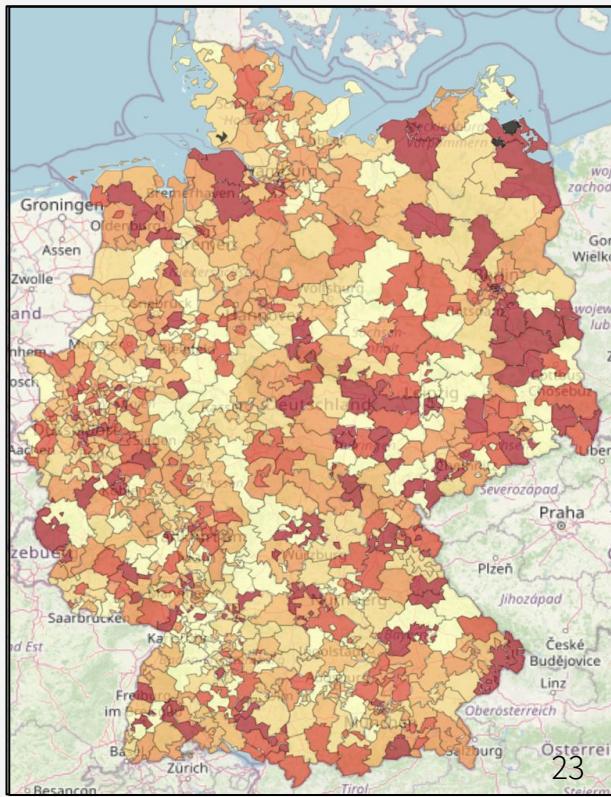
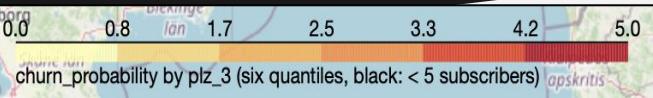
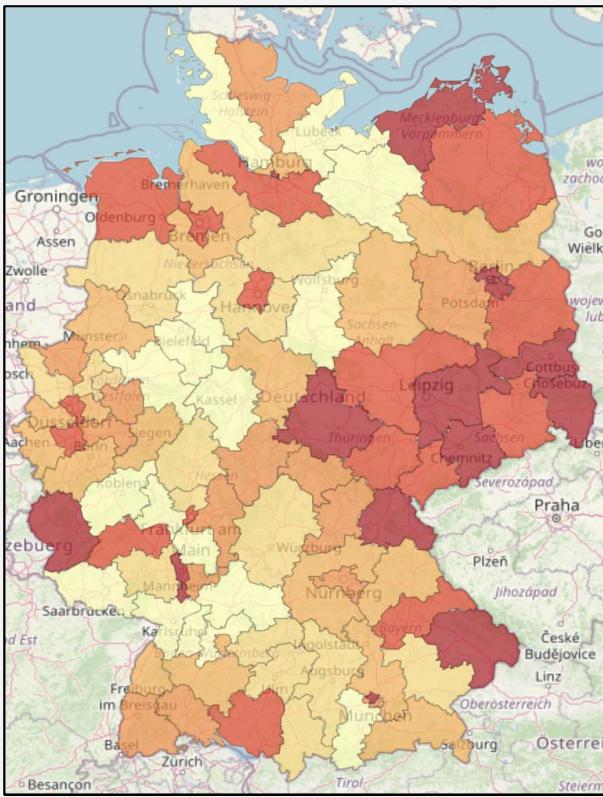
- Churn rate decreases with increasing reading time
- 12 months of reading is the major threshold towards loyal customers
- Smaller peaks in the churn rate, after two years, three years, etc.



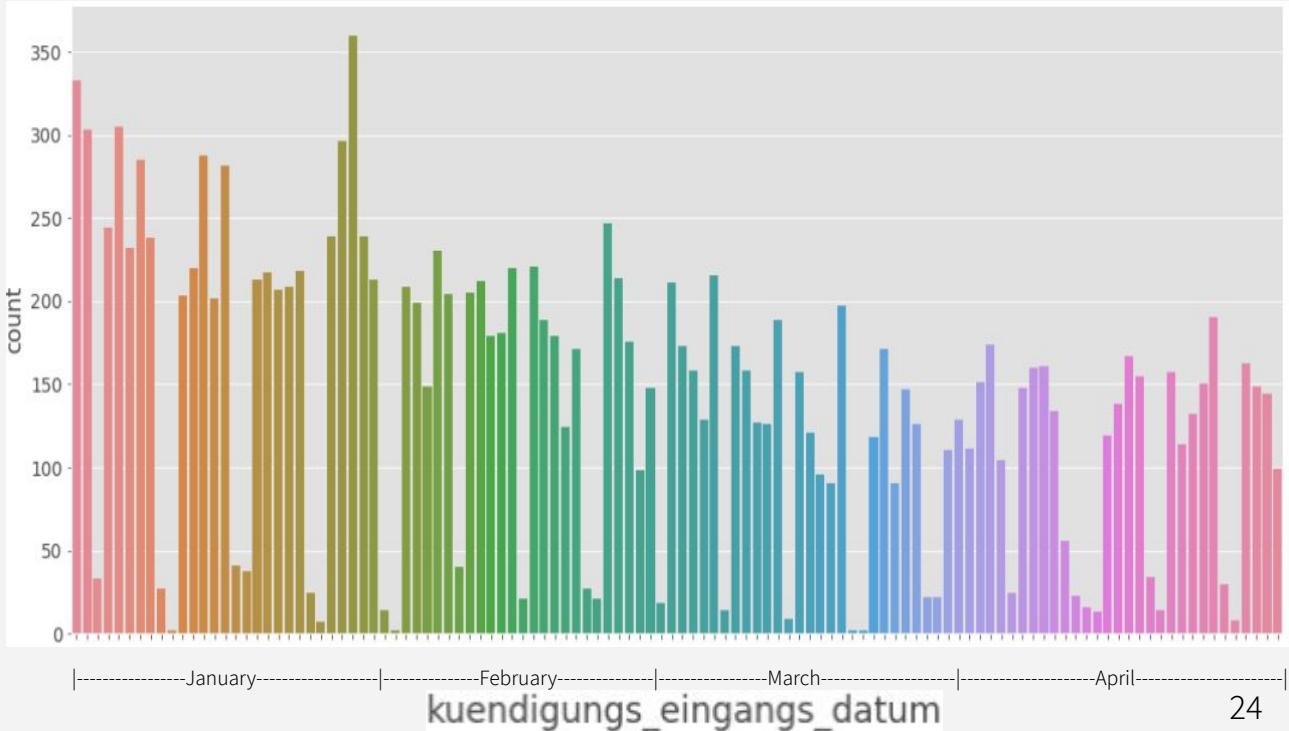
DIGITAL VS. PAPER VS. CHRIST & WELT



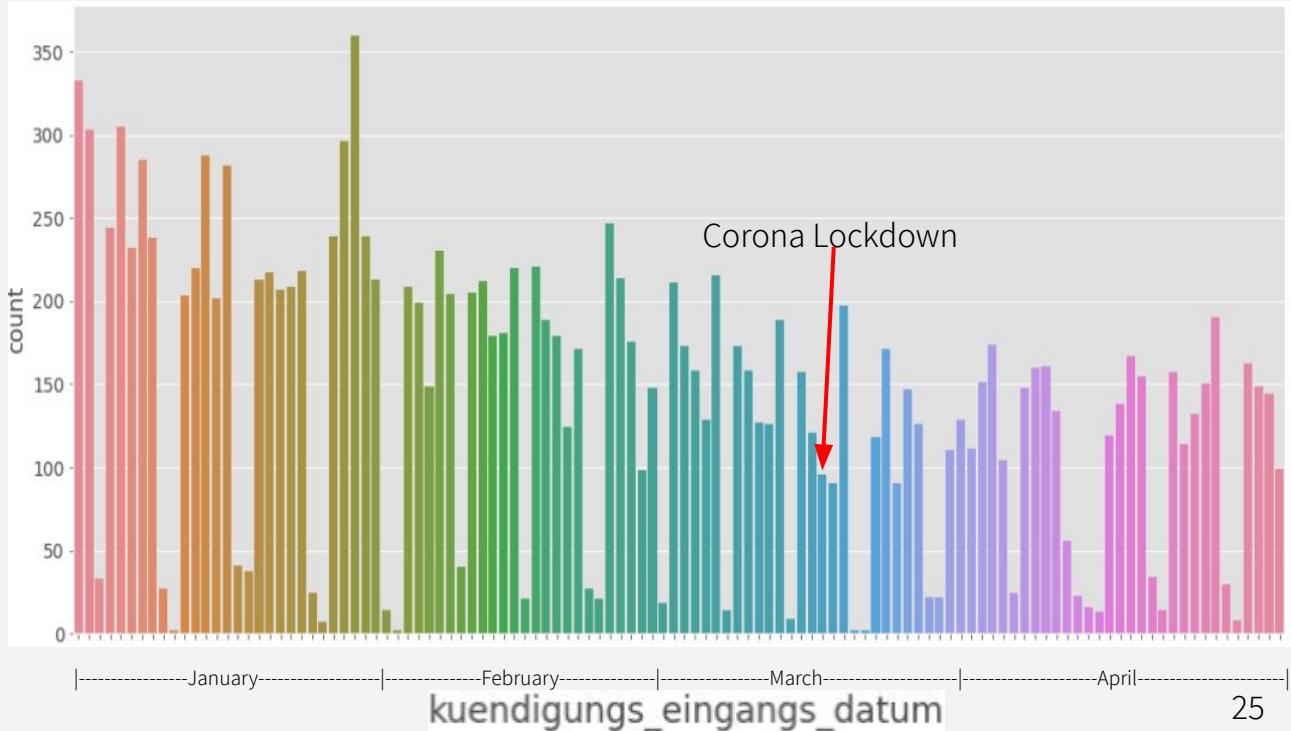
POSTAL CODE



DOES CORONA HAVE AN EFFECT ON THE CHURN RATE?

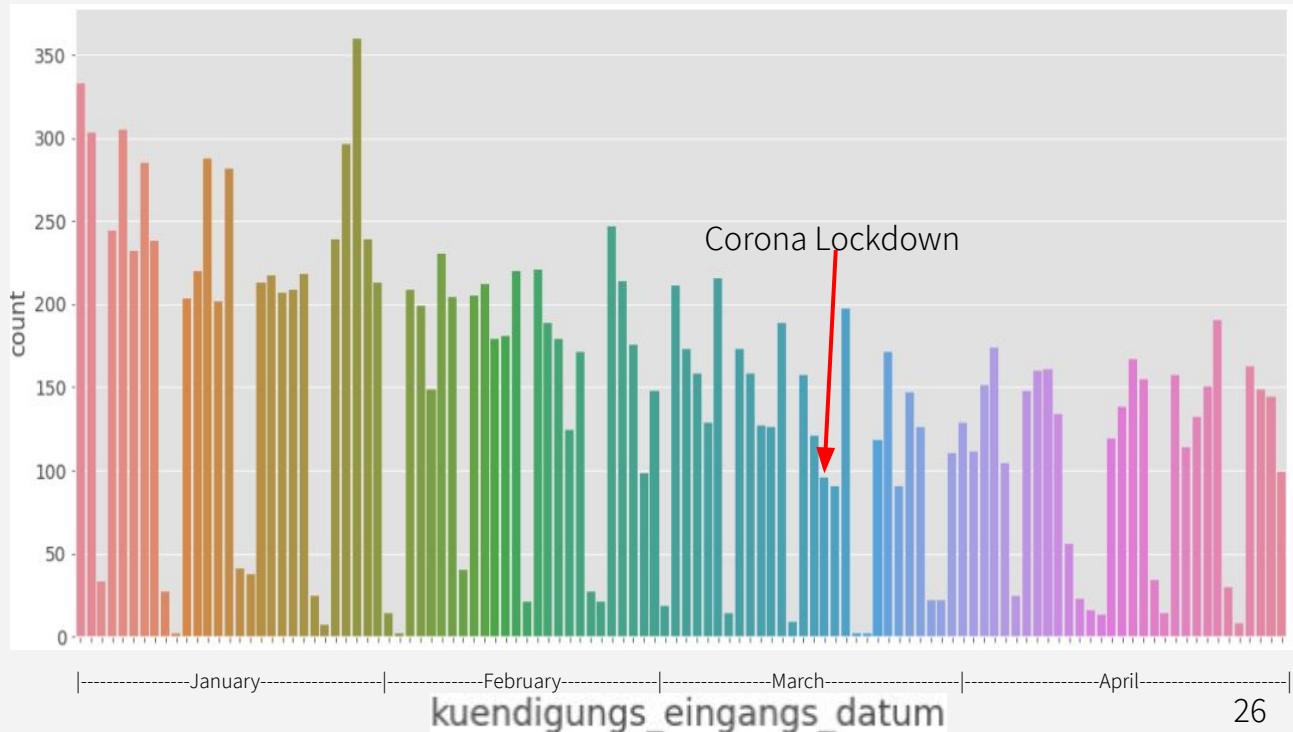


DOES CORONA HAVE AN EFFECT ON THE CHURN RATE?



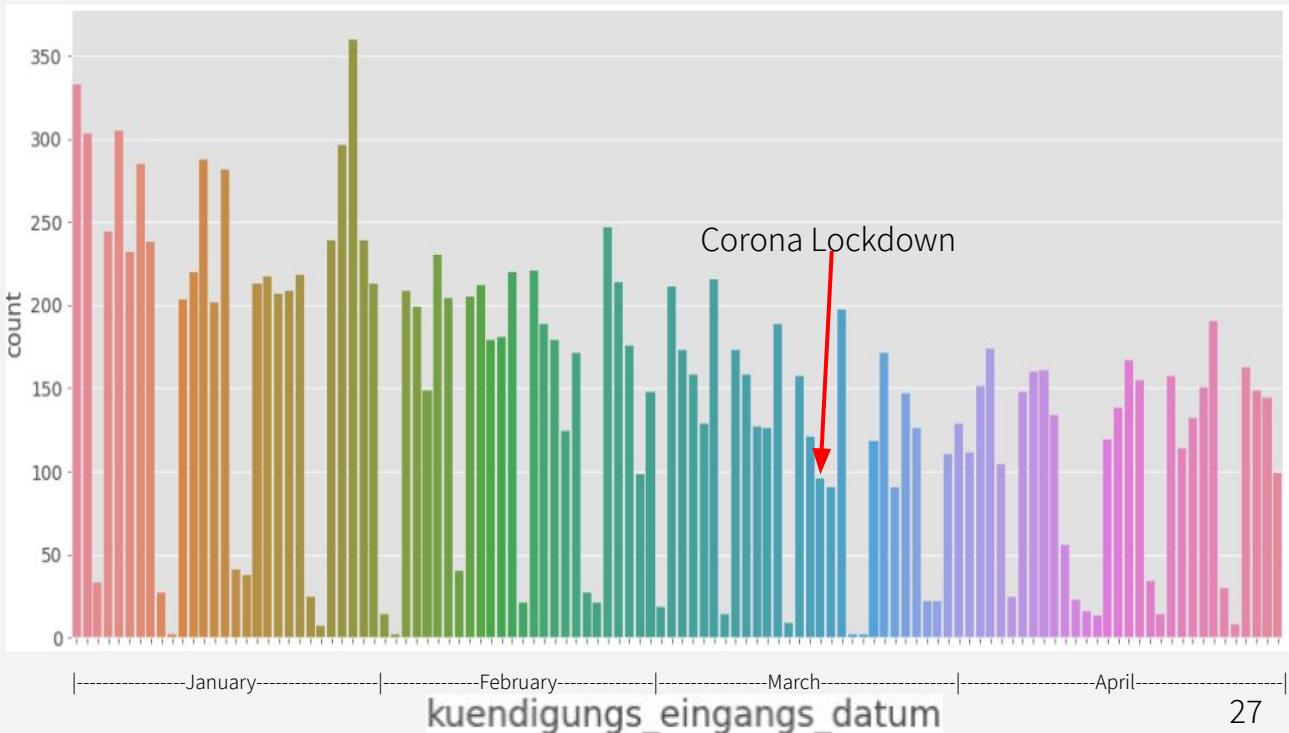
DOES CORONA HAVE AN EFFECT ON THE CHURN RATE?

- We do not have a comparable period, as we are only looking at one year



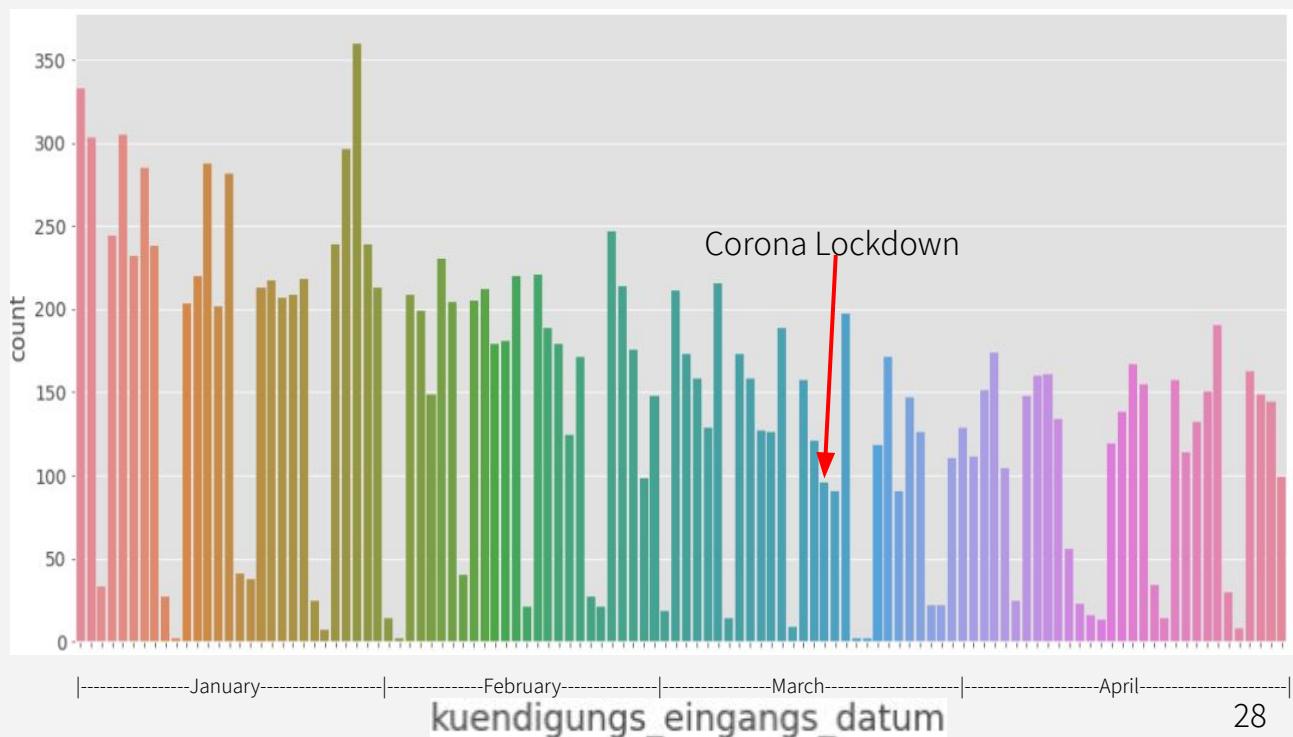
DOES CORONA HAVE AN EFFECT ON THE CHURN RATE?

- 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



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

THREE ALGORITHMS FOR FEATURE SELECTION

1. Correlation with churn

THREE ALGORITHMS FOR FEATURE SELECTION

1. Correlation with churn
2. SelectKBest from Scikit-Learn

THREE ALGORITHMS FOR FEATURE SELECTION

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

THREE ALGORITHMS FOR FEATURE SELECTION

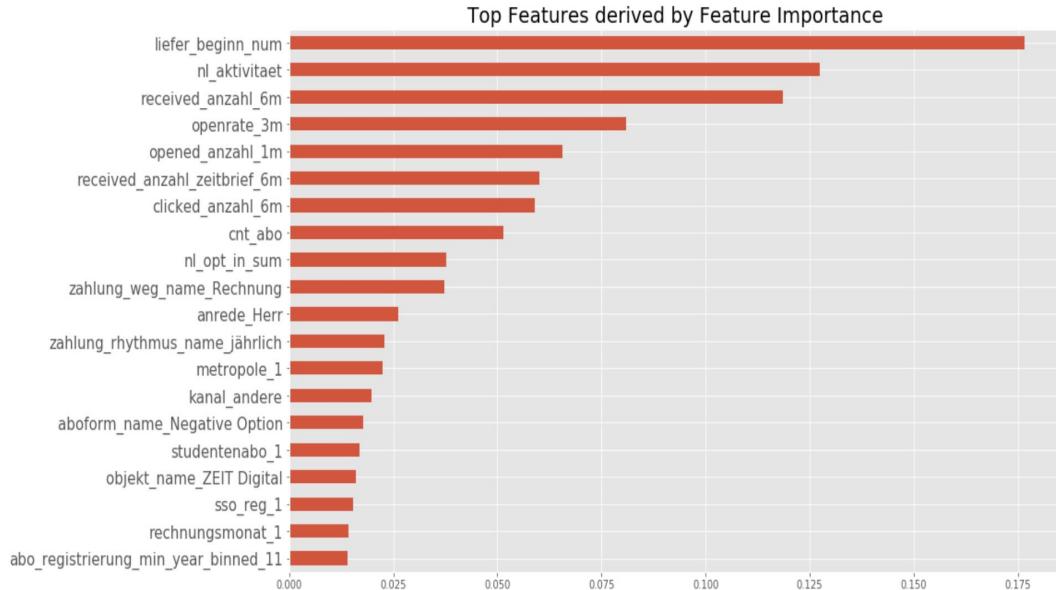
1. Correlation with churn
2. SelectKBest from Scikit-Learn
3. Feature importance from decision trees

→ Several different feature sets with 20 or 30 features each

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





04

MACHINE LEARNING MODELS

WHAT IS OUR GOAL?



WHAT IS OUR GOAL?



ACCURACY

How often has the model been right?

WHAT IS OUR GOAL?



ACCURACY

How often has the model been right?



RECALL

How many churns are detected ?

WHAT IS OUR GOAL?



PRECISION

Correctly predicted churns in relation to false alarms



ACCURACY

How often has the model been right?



RECALL

How many churns are detected ?

WHAT IS OUR GOAL?

Emphasis on precision, because false alarms may cost money



PRECISION

Correctly predicted churns in relation to false alarms



ACCURACY

How often has the model been right?



RECALL

How many churns are detected ?

WHAT IS OUR GOAL?

Emphasis on precision, because false alarms may cost money



PRECISION

Correctly predicted churns in relation to false alarms



ACCURACY

How often has the model been right?



RECALL

How many churns are detected ?



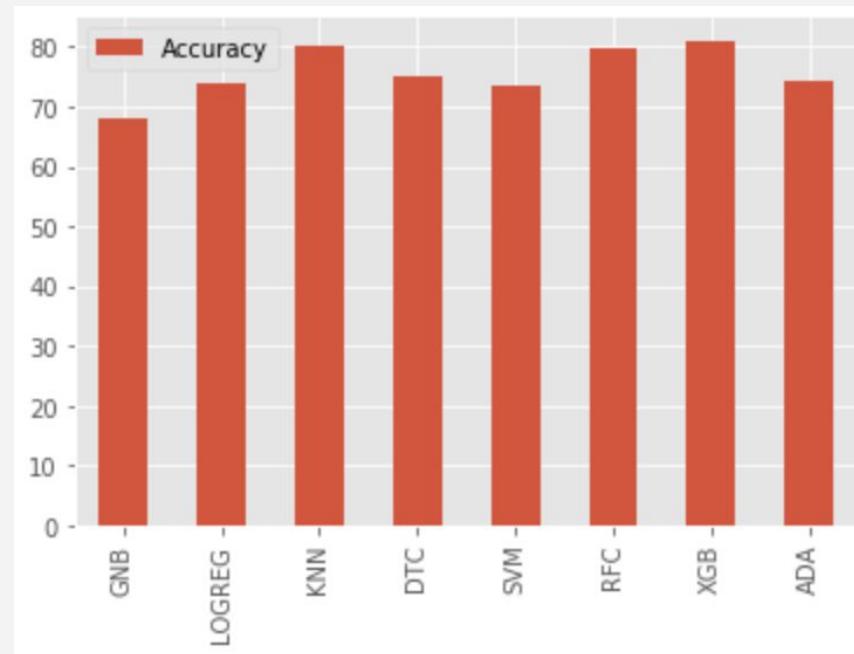
F-BETA-SCORE

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

MACHINE LEARNING MODELS

USED MODELS:

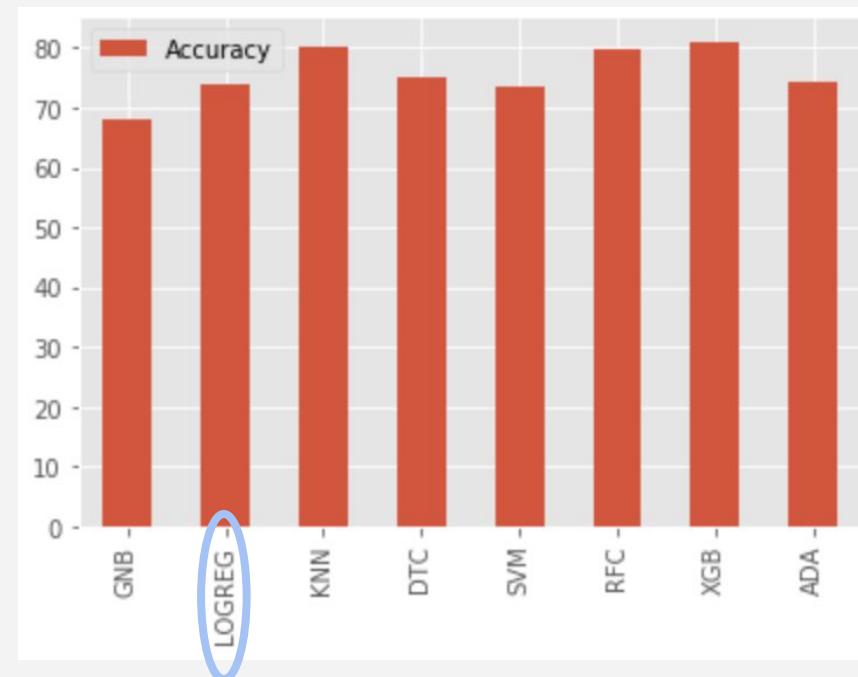
1. Gaussian Naive Bayes
2. Logistic Regression
3. K Nearest Neighbors
4. Decision Tree
5. Support Vector Machines
6. Random Forest
7. XGBoost
8. AdaBoost



MACHINE LEARNING MODELS

USED MODELS:

- Baseline Model
- 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

USED 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

Baseline Model

Best Models



MACHINE LEARNING MODELS

USED 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

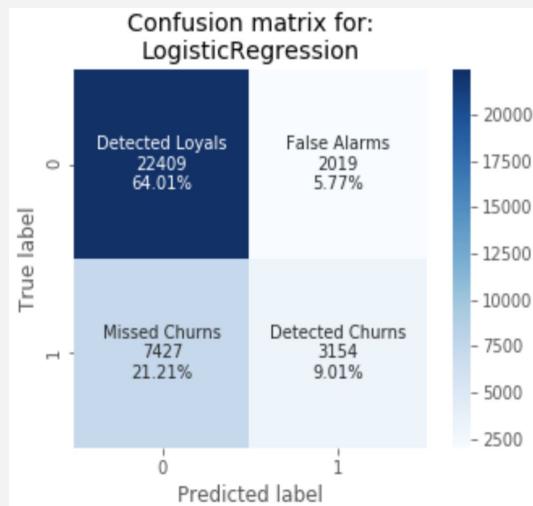
Baseline Model

Best Models



BEST MODELS VS BASELINE MODEL

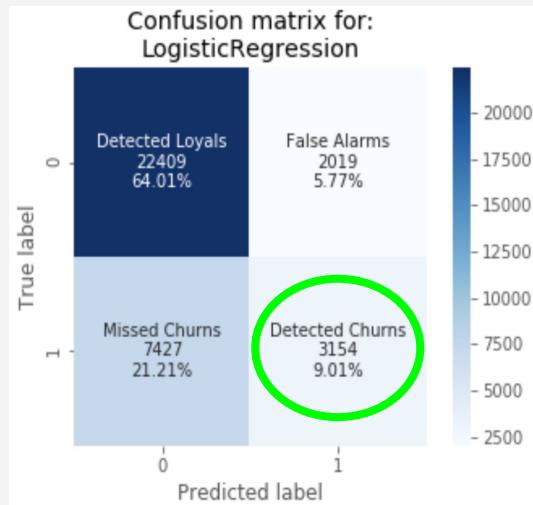
LOGISTIC REGRESSION BASELINE MODEL



ACCURACY:
RECALL:
PRECISION:

BEST MODELS VS BASELINE MODEL

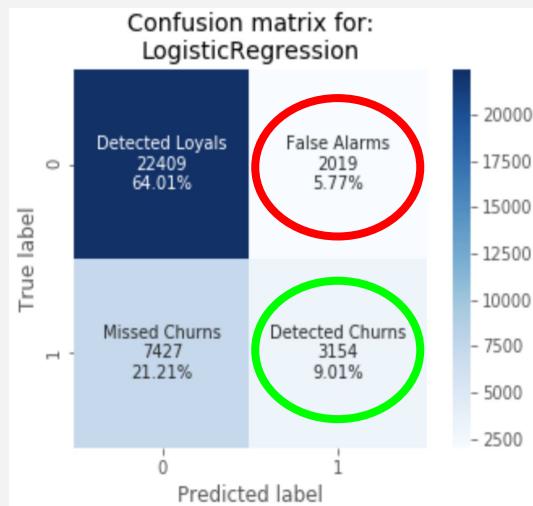
LOGISTIC REGRESSION BASELINE MODEL



ACCURACY:
RECALL:
PRECISION:

BEST MODELS VS BASELINE MODEL

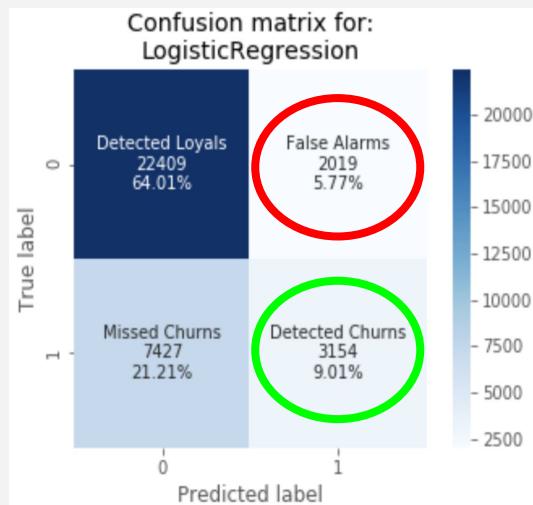
LOGISTIC REGRESSION BASELINE MODEL



ACCURACY:
RECALL:
PRECISION:

BEST MODELS VS BASELINE MODEL

LOGISTIC REGRESSION BASELINE MODEL

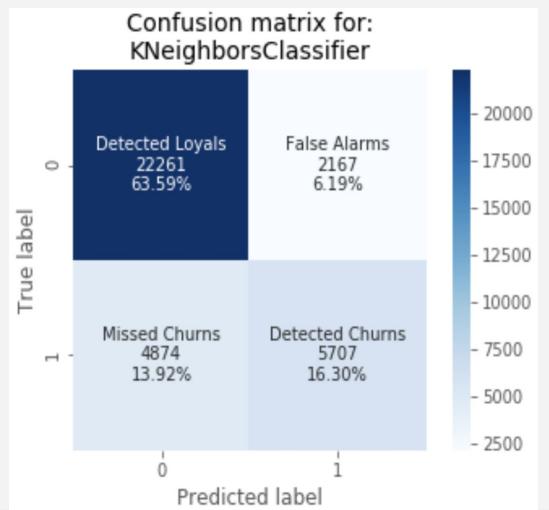


ACCURACY:
RECALL:
PRECISION:

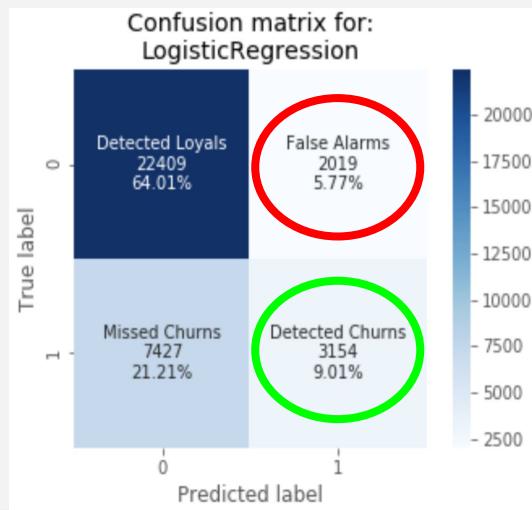
73.0%
29.8%
61.0%

BEST MODELS VS BASELINE MODEL

KNN



LOGISTIC REGRESSION BASELINE MODEL

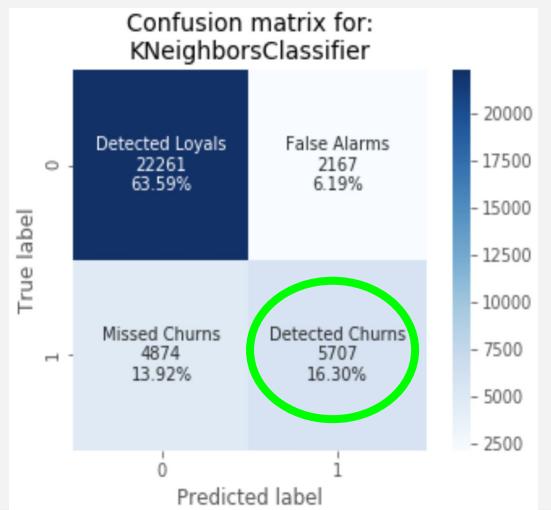


ACCURACY:
RECALL:
PRECISION:

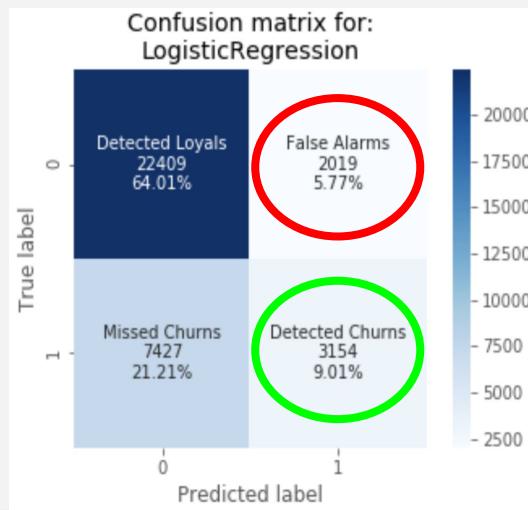
73.0%
29.8%
61.0%

BEST MODELS VS BASELINE MODEL

KNN



LOGISTIC REGRESSION BASELINE MODEL

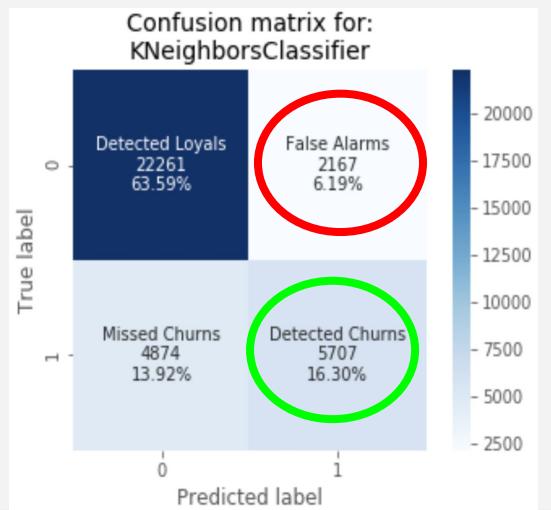


ACCURACY:
RECALL:
PRECISION:

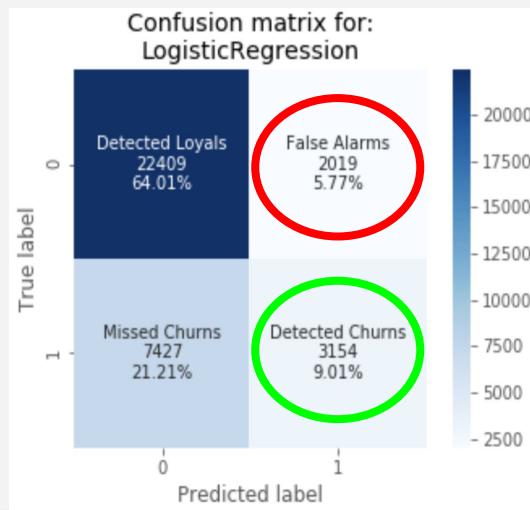
73.0%
29.8%
61.0%

BEST MODELS VS BASELINE MODEL

KNN



LOGISTIC REGRESSION BASELINE MODEL

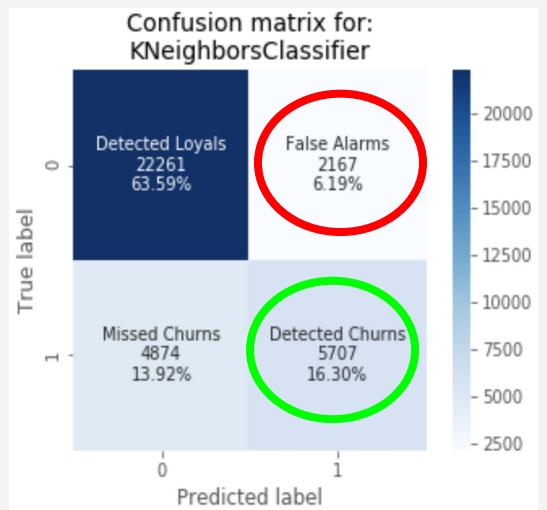


ACCURACY:
RECALL:
PRECISION:

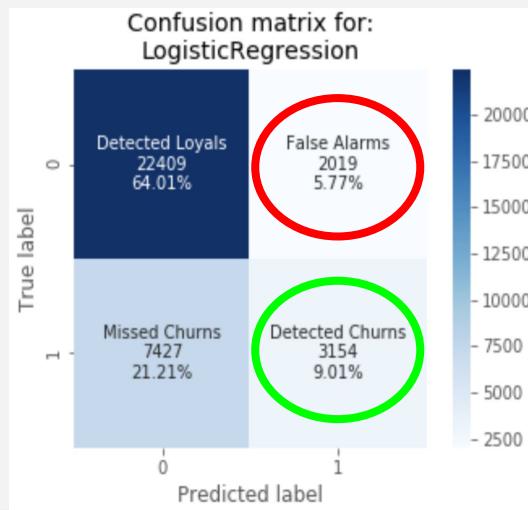
73.0%
29.8%
61.0%

BEST MODELS VS BASELINE MODEL

KNN



LOGISTIC REGRESSION BASELINE MODEL

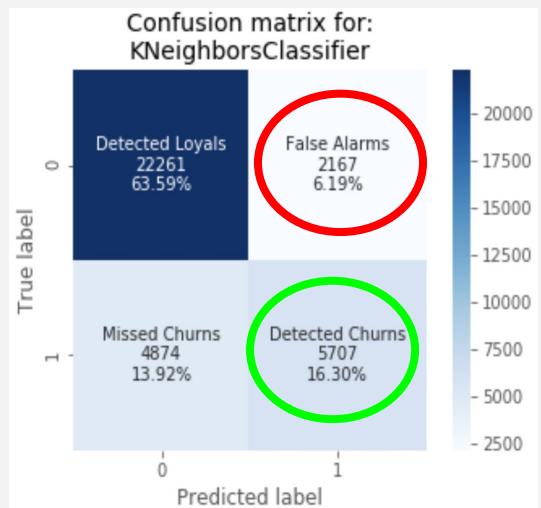


ACCURACY: 79.9%
RECALL: 53.9%
PRECISION: 72.5%

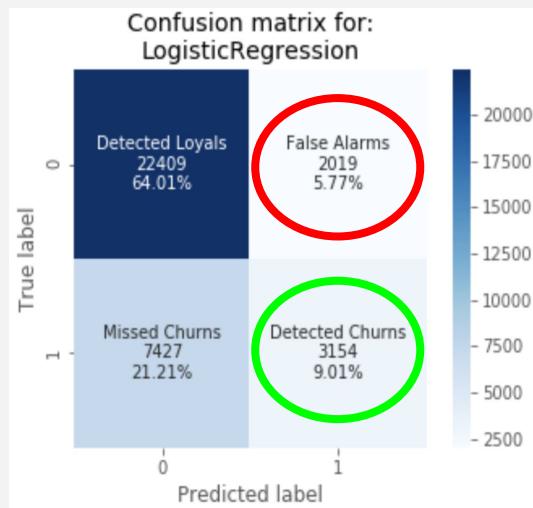
73.0%
29.8%
61.0%

BEST MODELS VS BASELINE MODEL

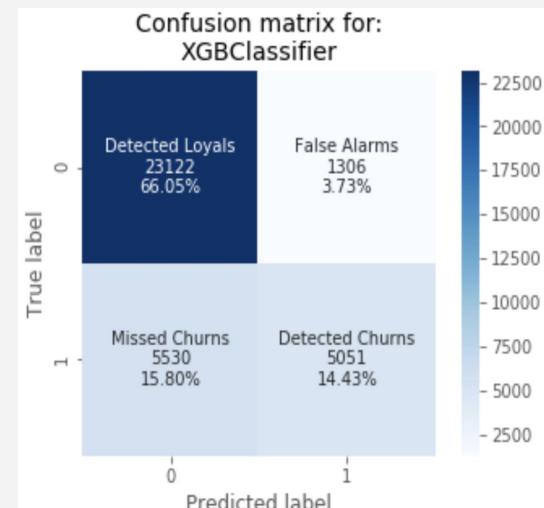
KNN



LOGISTIC REGRESSION BASELINE MODEL



XGBoost

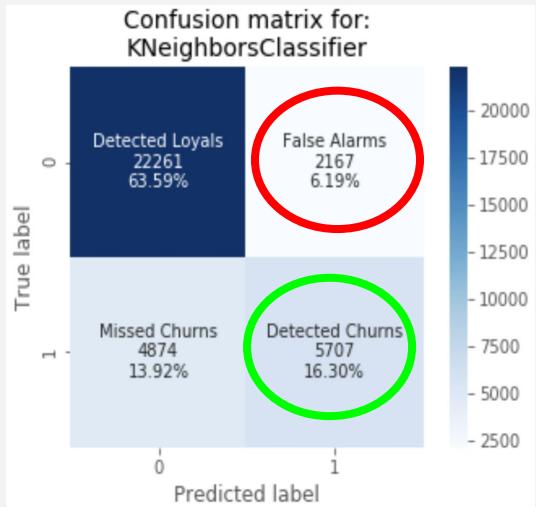


ACCURACY: 79.9%
RECALL: 53.9%
PRECISION: 72.5%

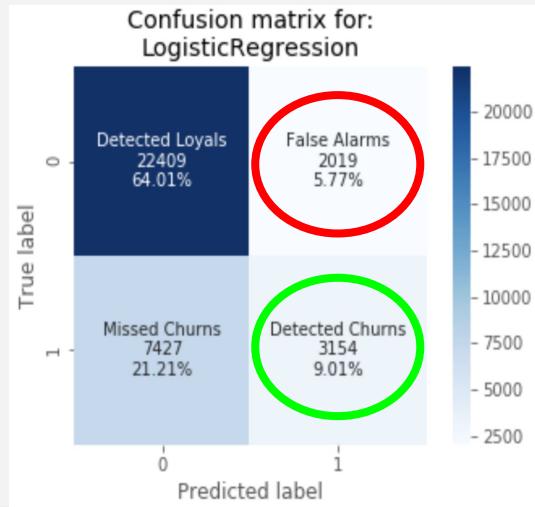
73.0%
29.8%
61.0%

BEST MODELS VS BASELINE MODEL

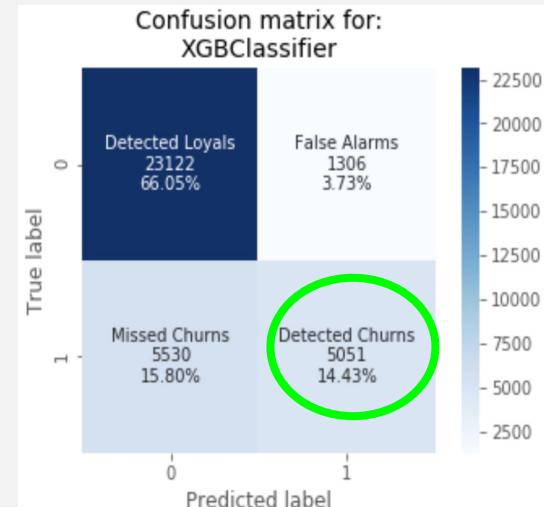
KNN



LOGISTIC REGRESSION BASELINE MODEL



XGBoost

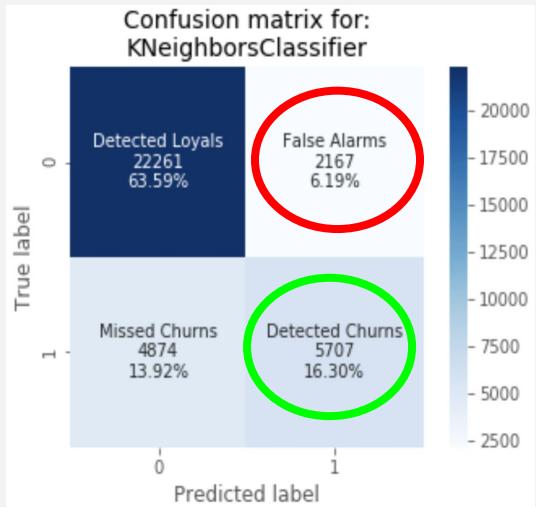


ACCURACY: 79.9%
RECALL: 53.9%
PRECISION: 72.5%

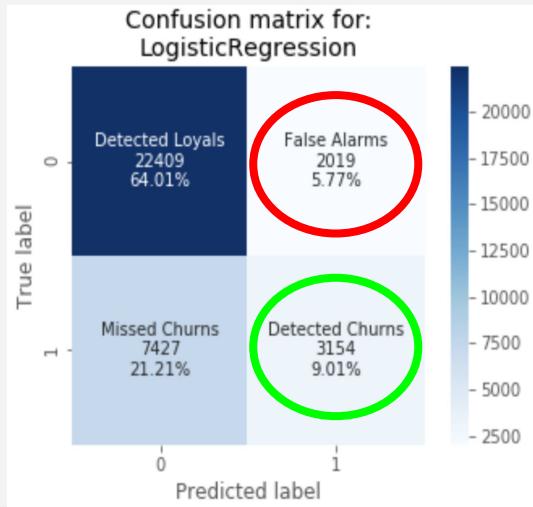
73.0%
29.8%
61.0%

BEST MODELS VS BASELINE MODEL

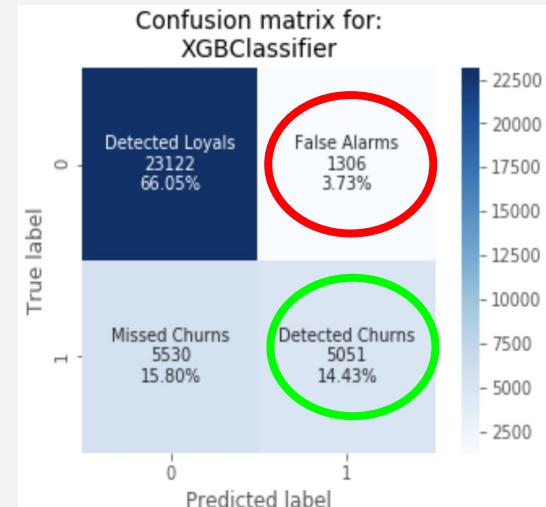
KNN



LOGISTIC REGRESSION BASELINE MODEL



XGBoost

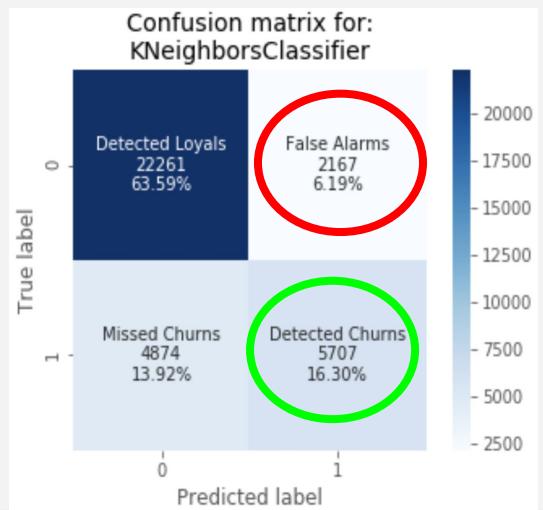


ACCURACY: 79.9%
RECALL: 53.9%
PRECISION: 72.5%

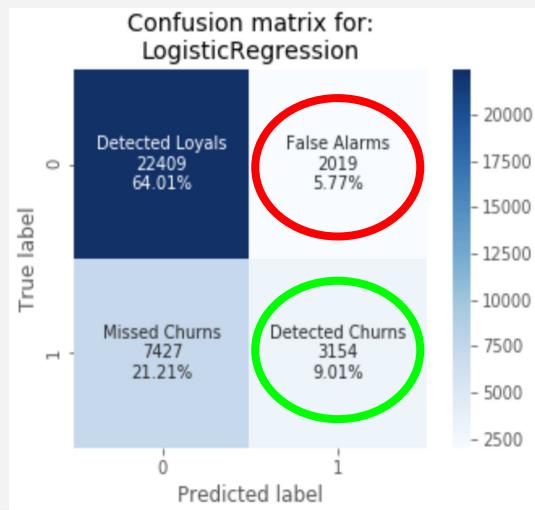
73.0%
29.8%
61.0%

BEST MODELS VS BASELINE MODEL

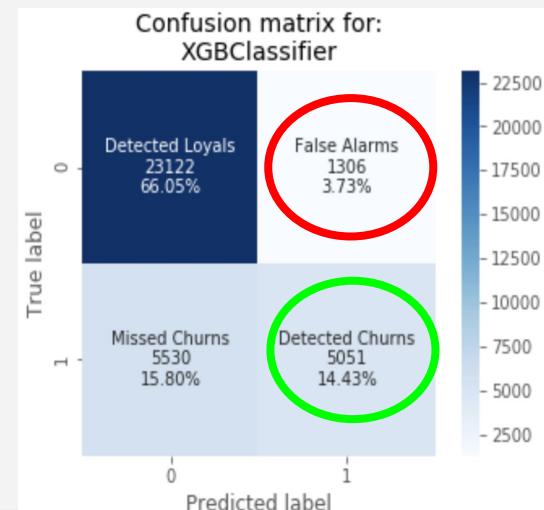
KNN



LOGISTIC REGRESSION BASELINE MODEL



XGBoost

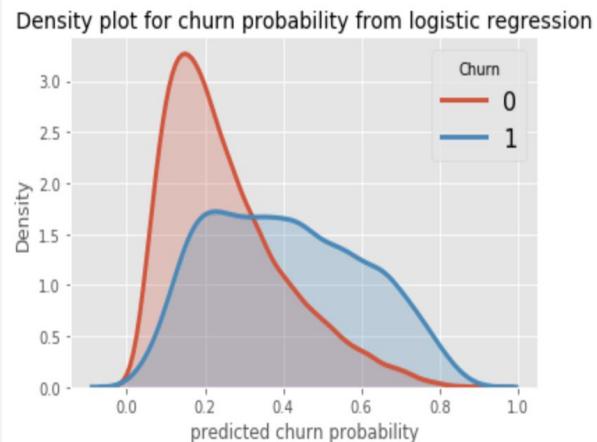


ACCURACY: 79.9%
RECALL: 53.9%
PRECISION: 72.5%

73.0%
29.8%
61.0%

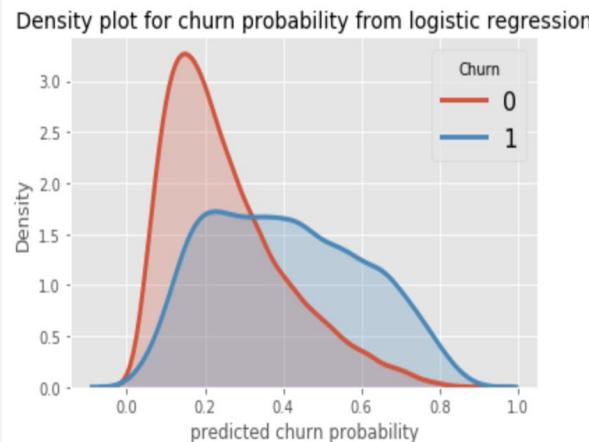
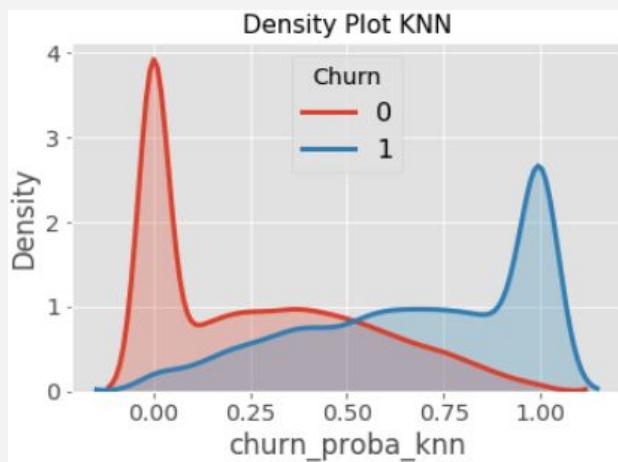
80.5%
47.7%
79.5%

SELECTIVITY OF THE MODELS



Base model (Logistic Regression)

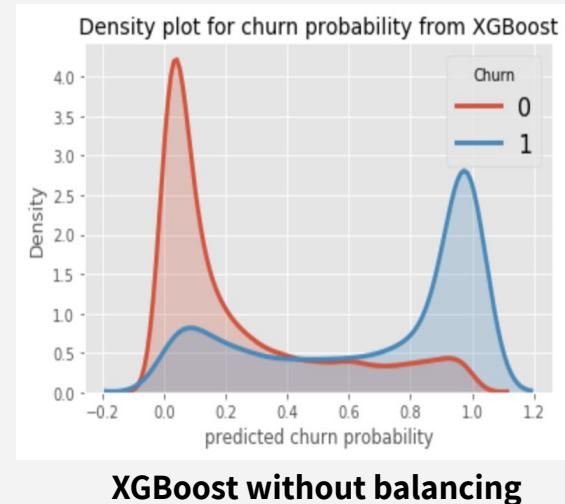
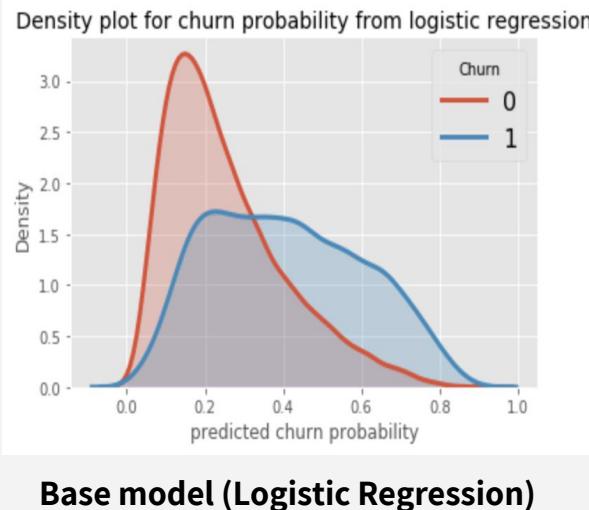
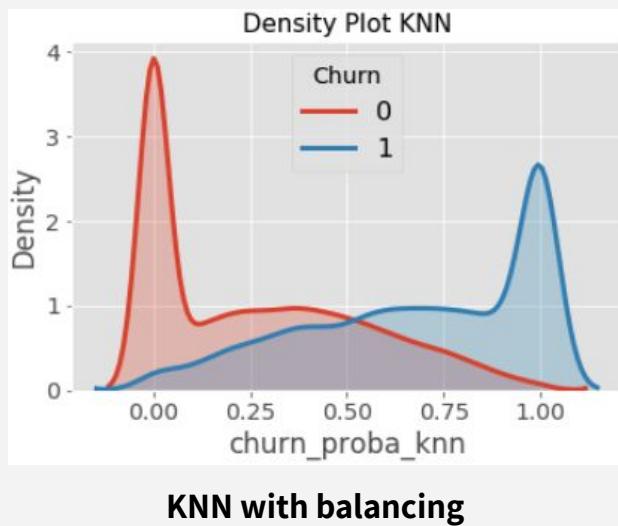
SELECTIVITY OF THE MODELS



Base model (Logistic Regression)

KNN with balancing

SELECTIVITY OF THE MODELS



SUMMARY

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



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.



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

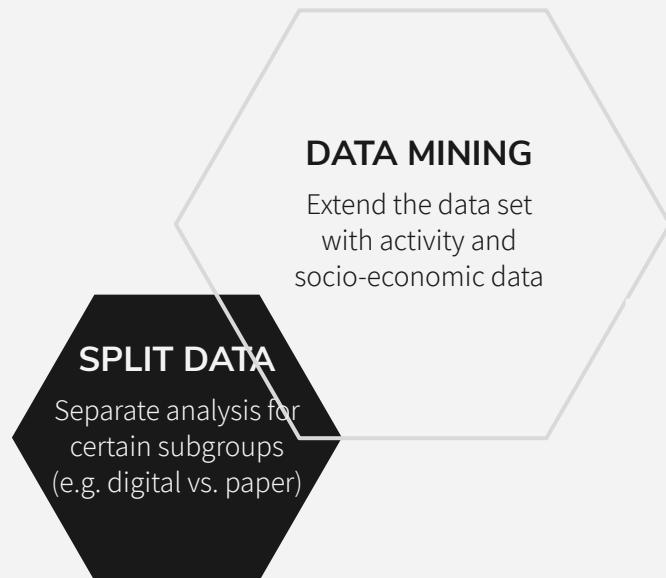


FUTURE WORK

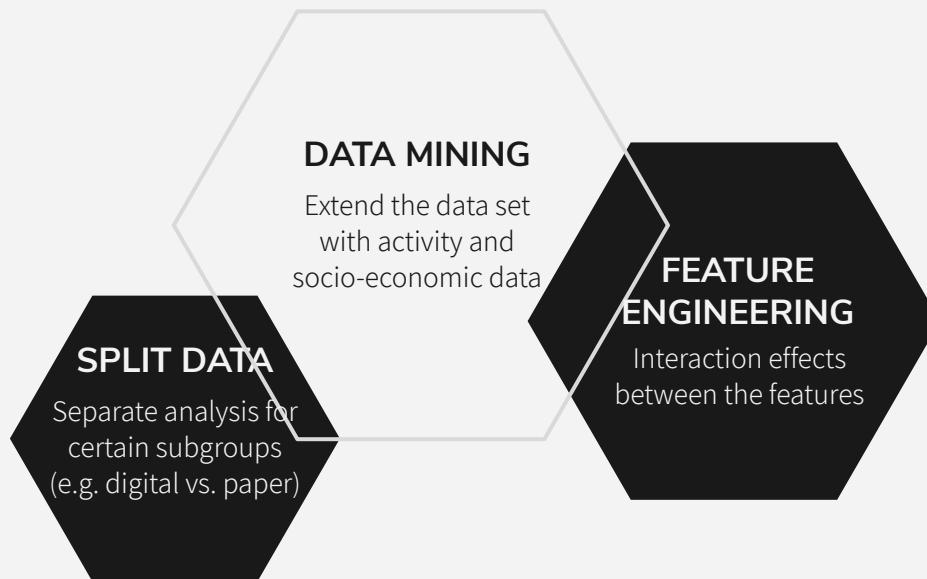
SPLIT DATA

Separate analysis for
certain subgroups
(e.g. digital vs. paper)

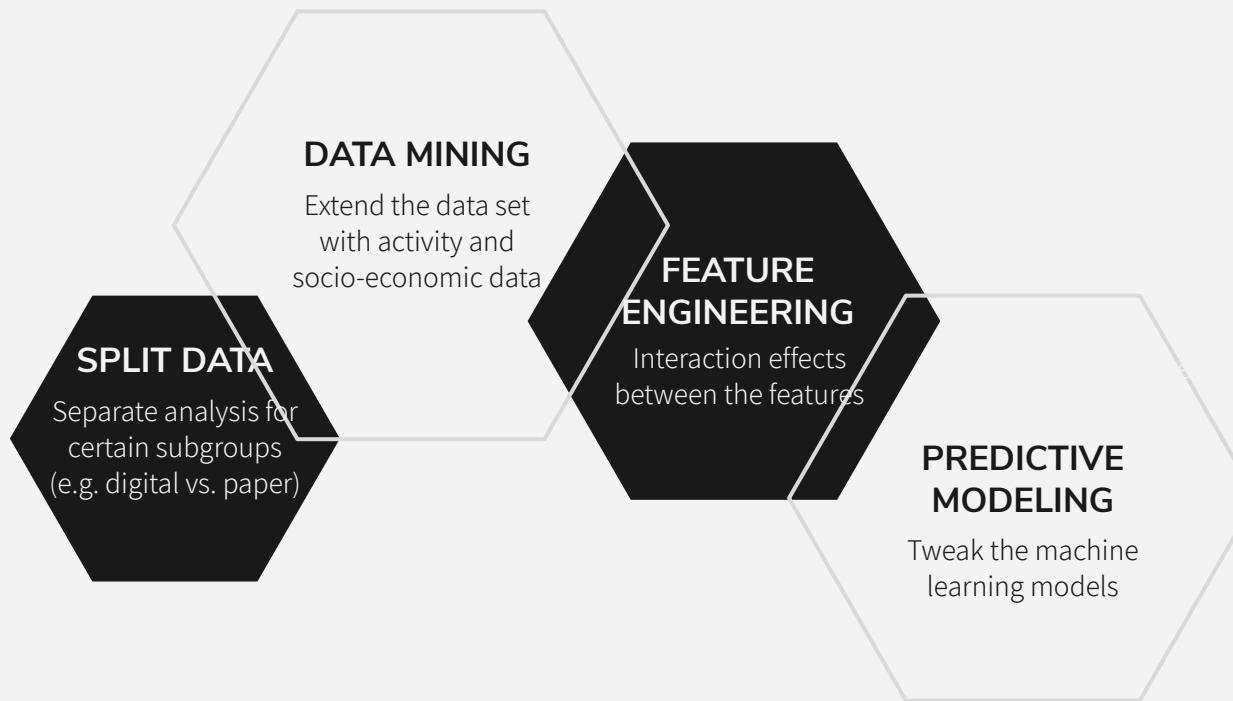
FUTURE WORK



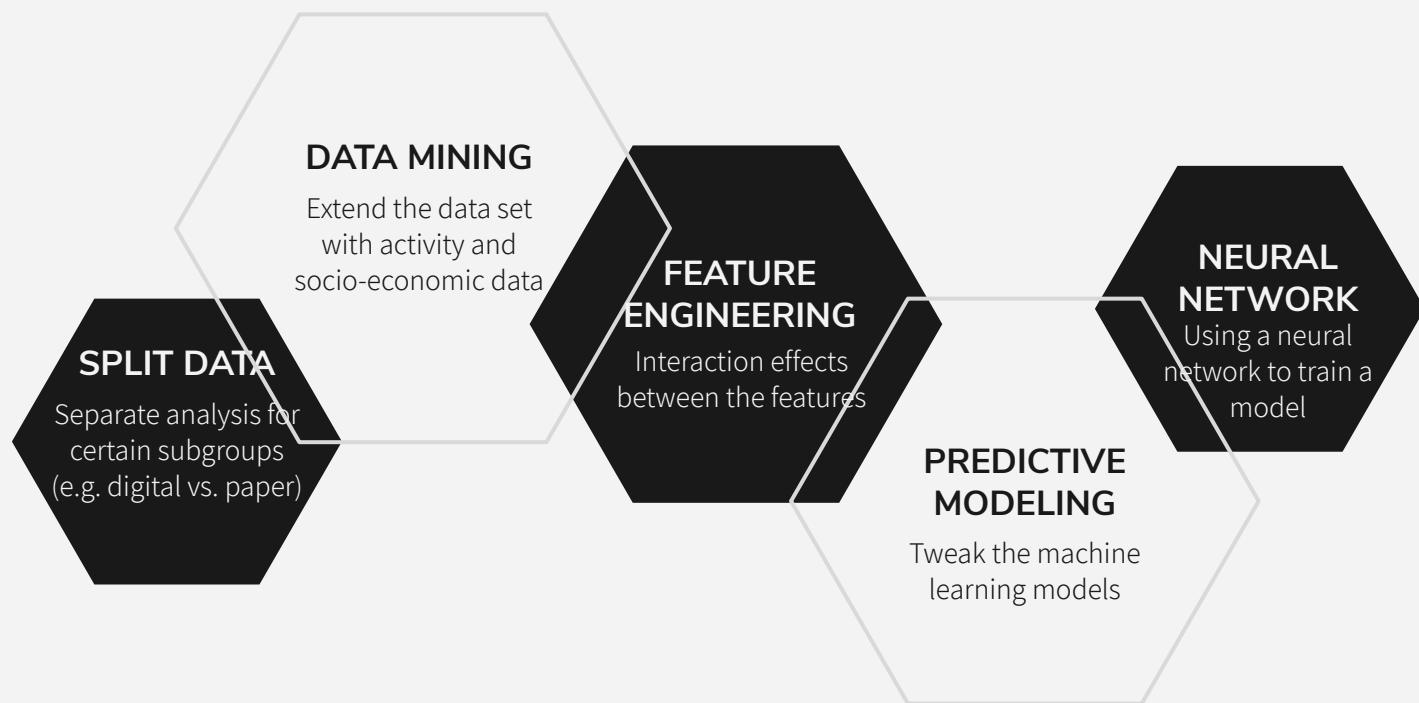
FUTURE WORK



FUTURE WORK



FUTURE WORK





THANKS

Does anyone have any questions?

fuehrer_sebastian at web.de

<https://github.com/booty7>

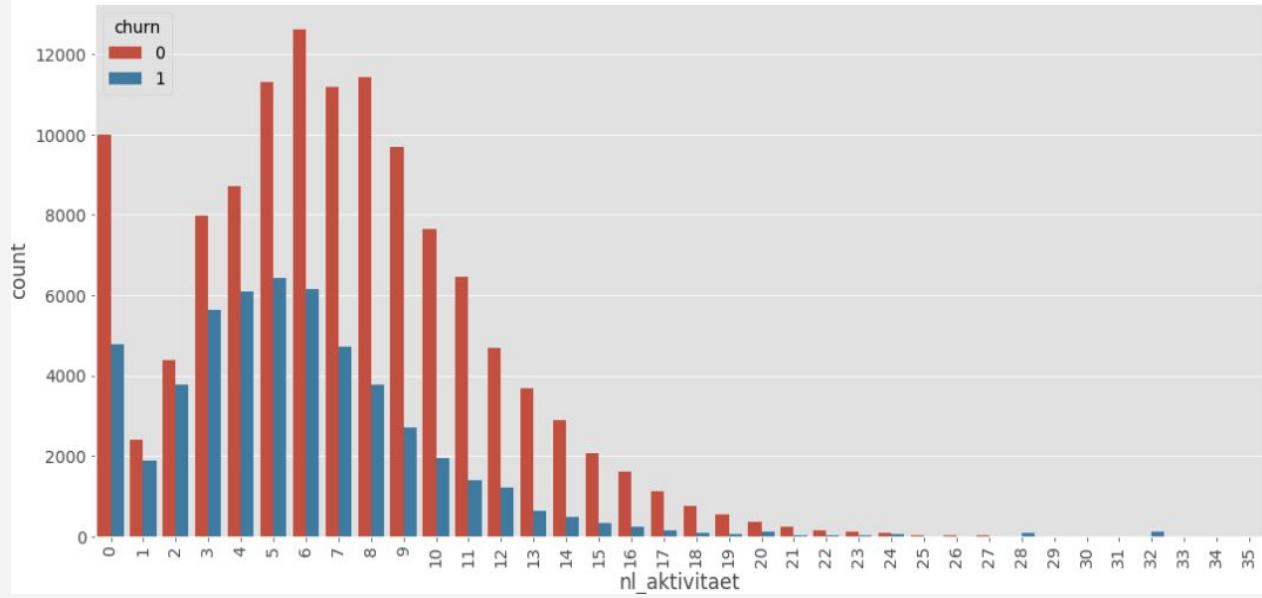
zagermannmarco at gmail.com

<https://github.com/mz2357>

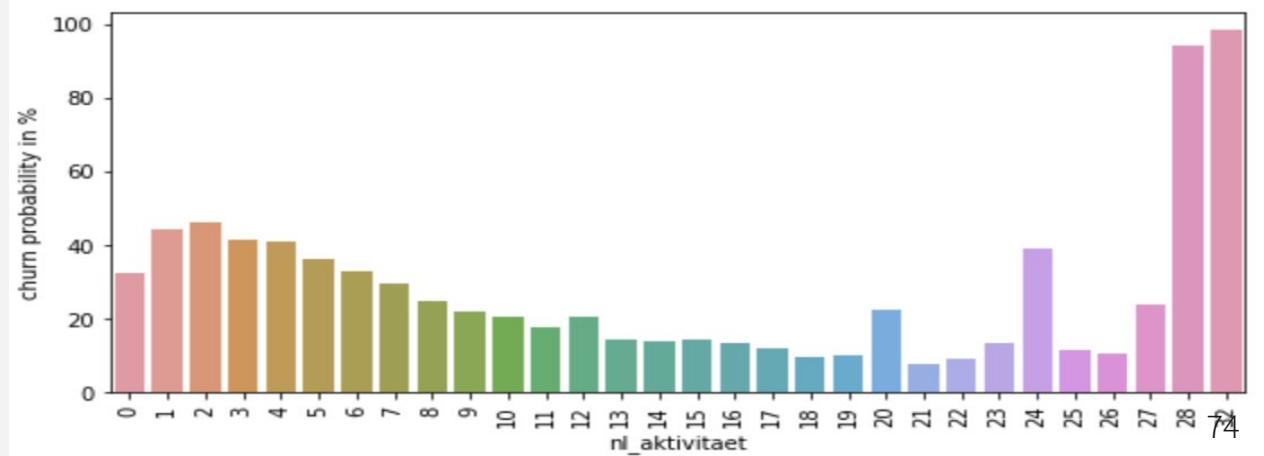
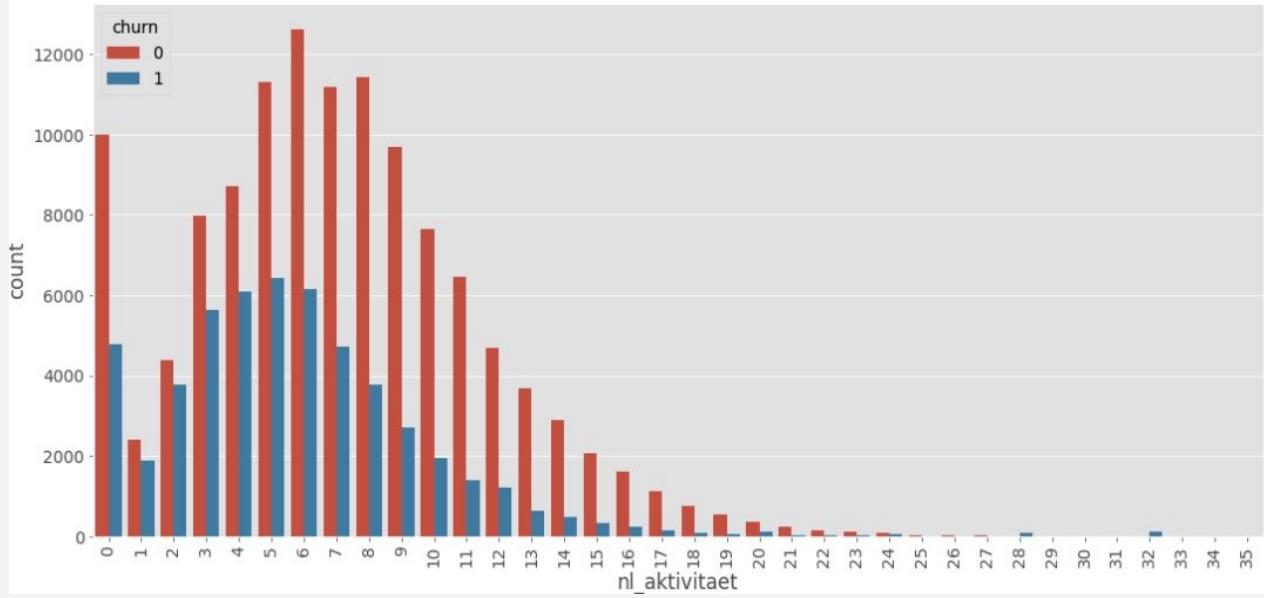




NEWSLETTER ACTIVITY

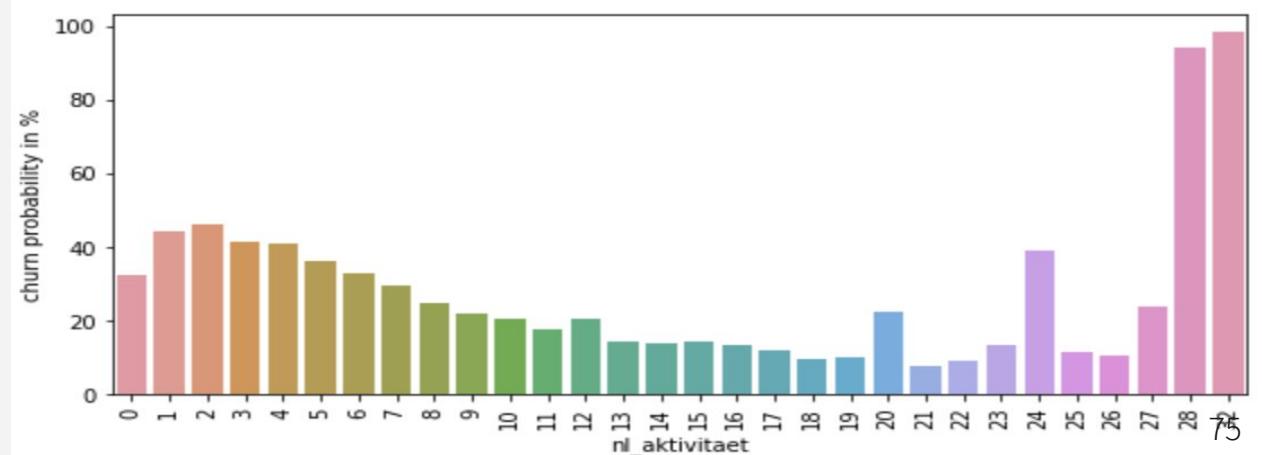
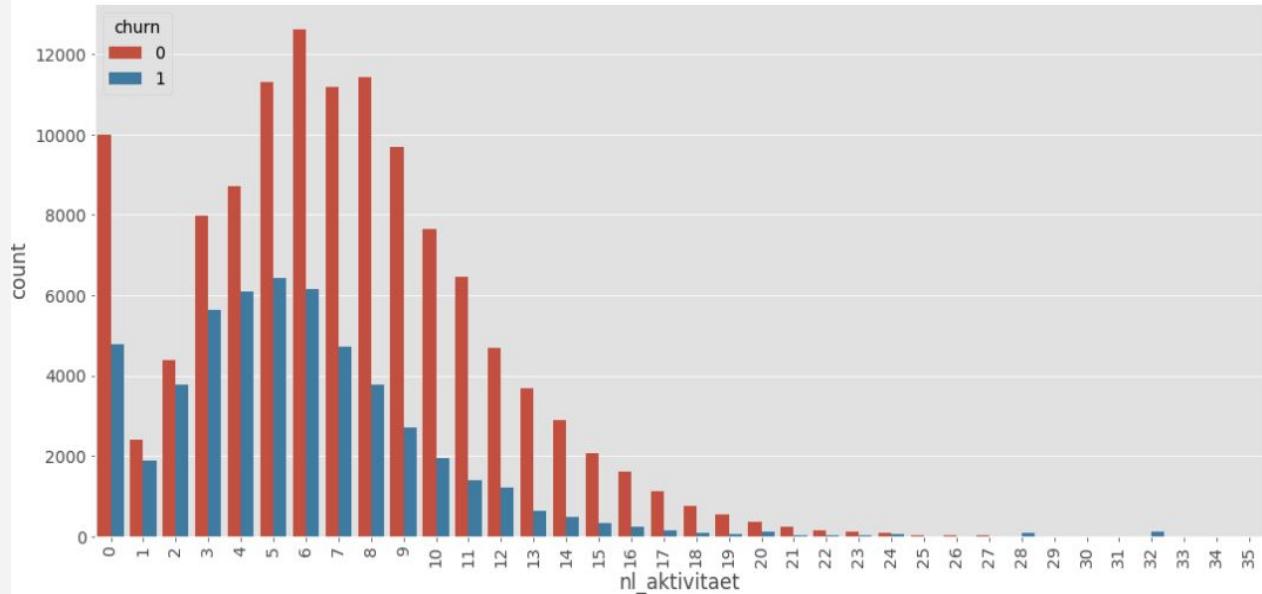


NEWSLETTER ACTIVITY



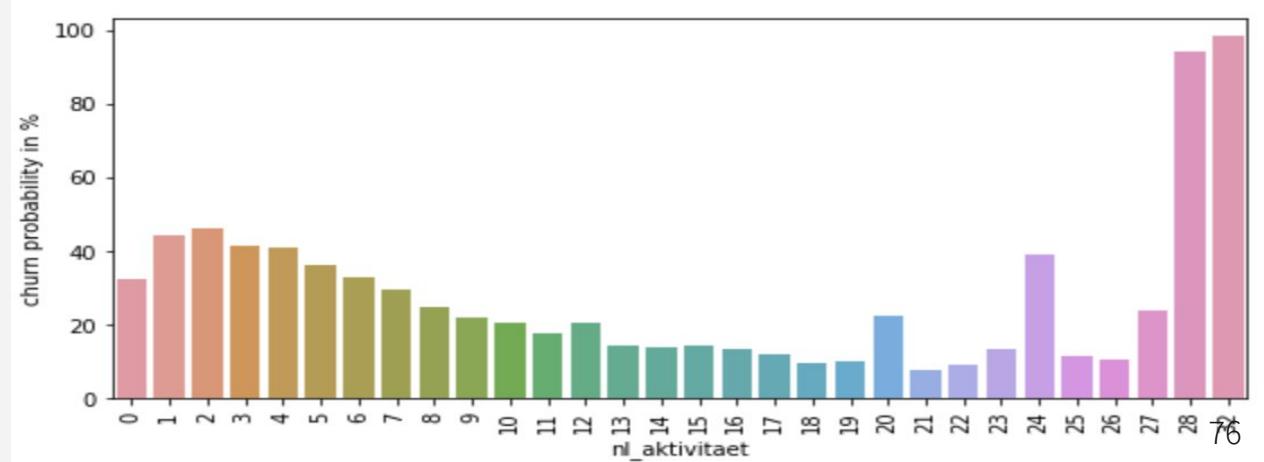
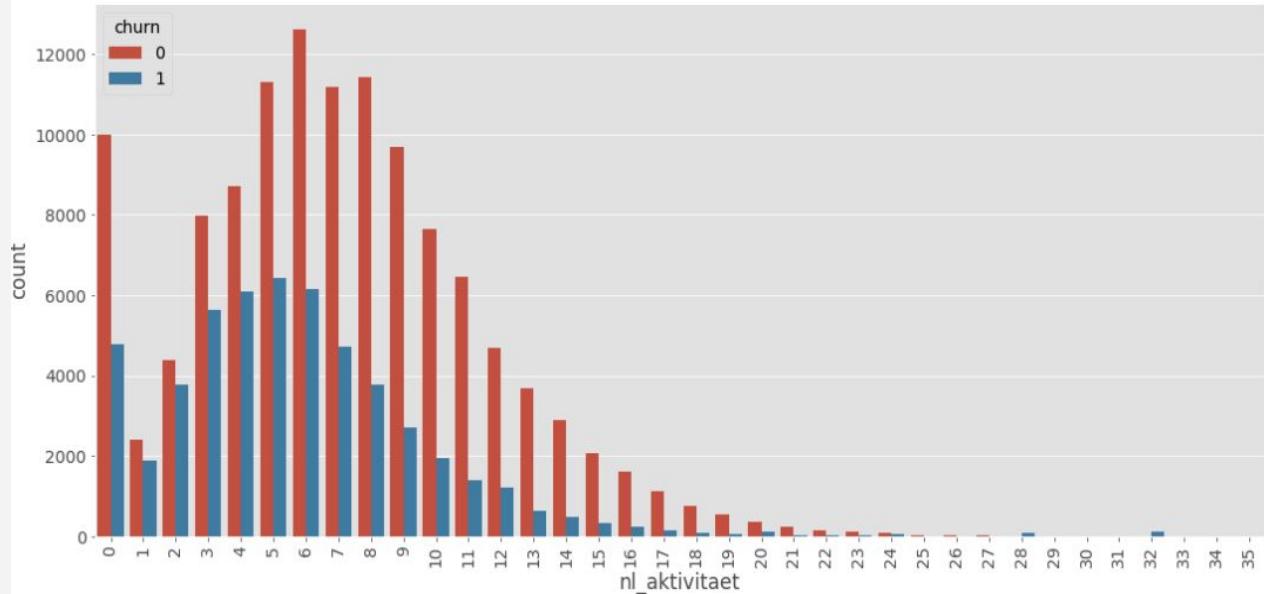
NEWSLETTER ACTIVITY

- Newsletter activity is a strong predictor

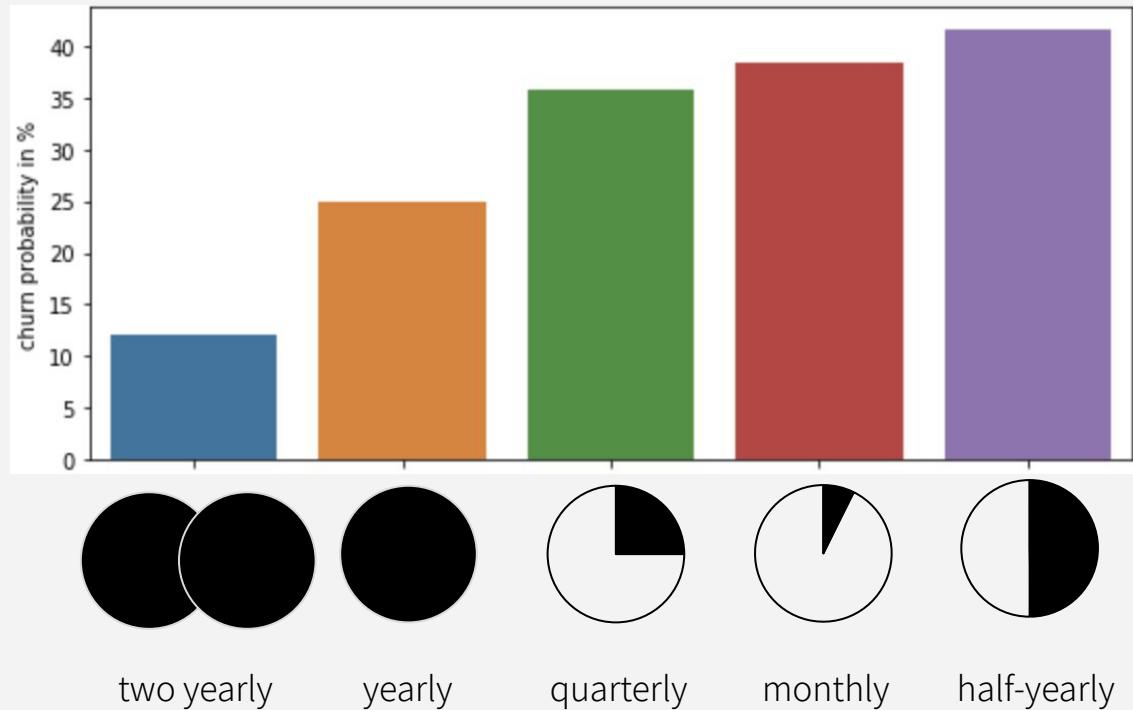


NEWSLETTER ACTIVITY

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



RHYTHM OF PAYMENT



Machine Learning Models optimized via Grid Search



K Nearest Neighbors

Best Model



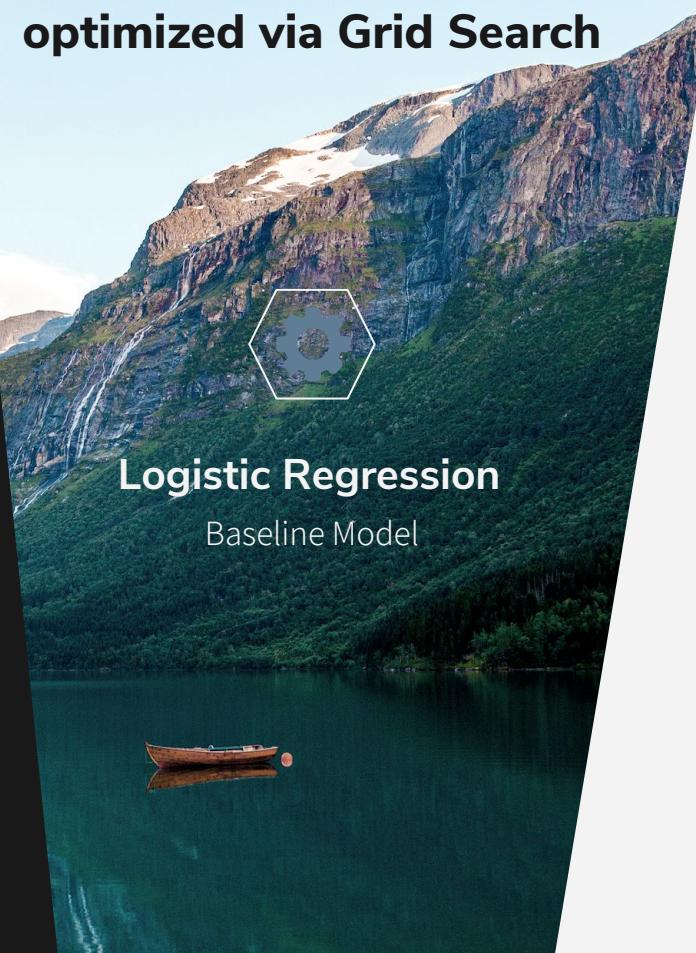
Logistic Regression

Baseline Model



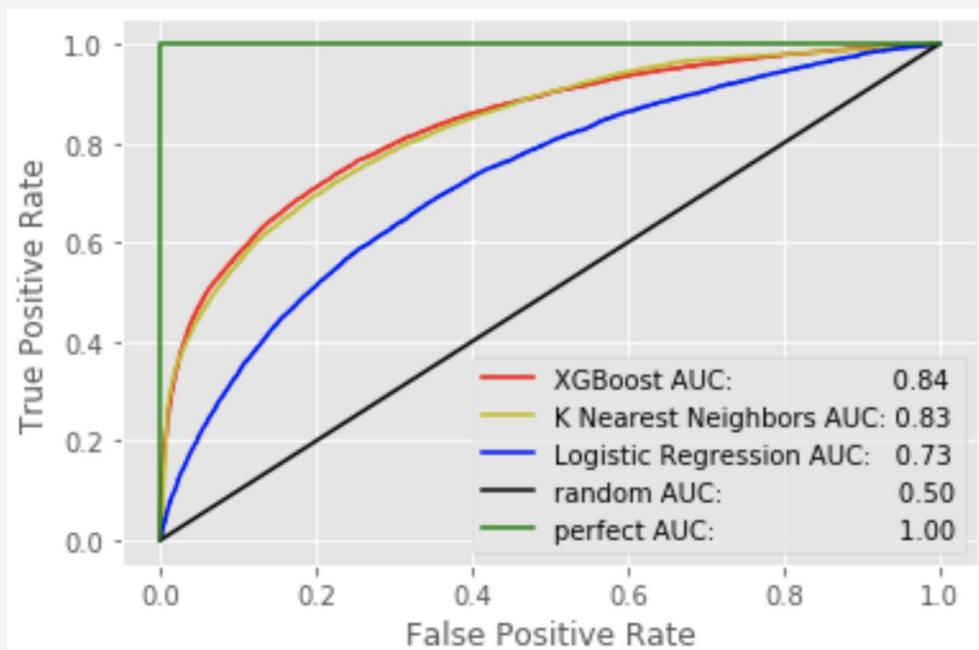
XGBoost

Best Model

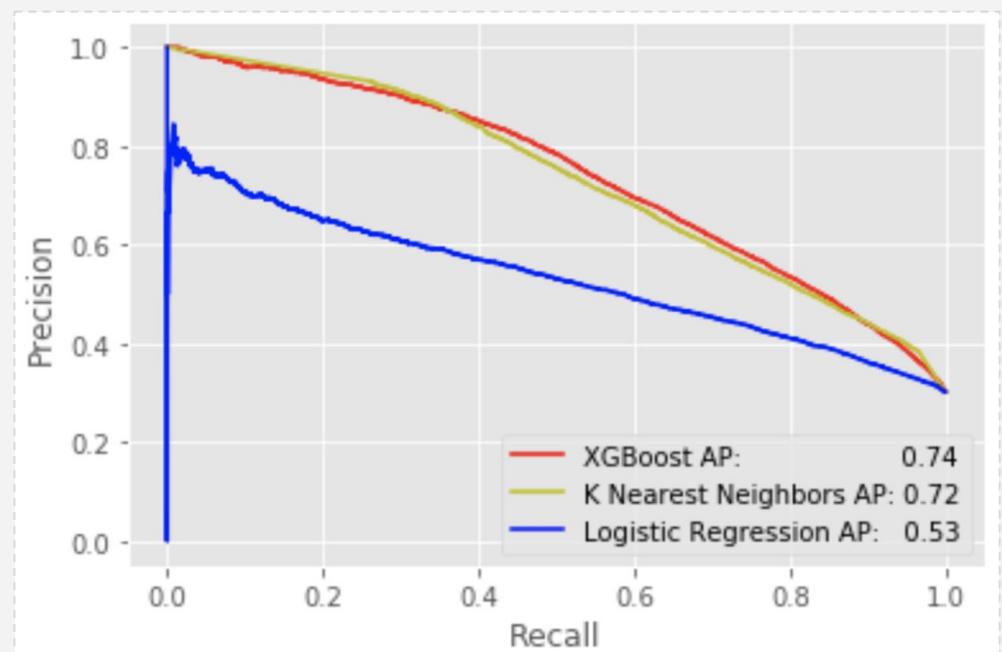


COMPARISON OF ROC AUC SCORE

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

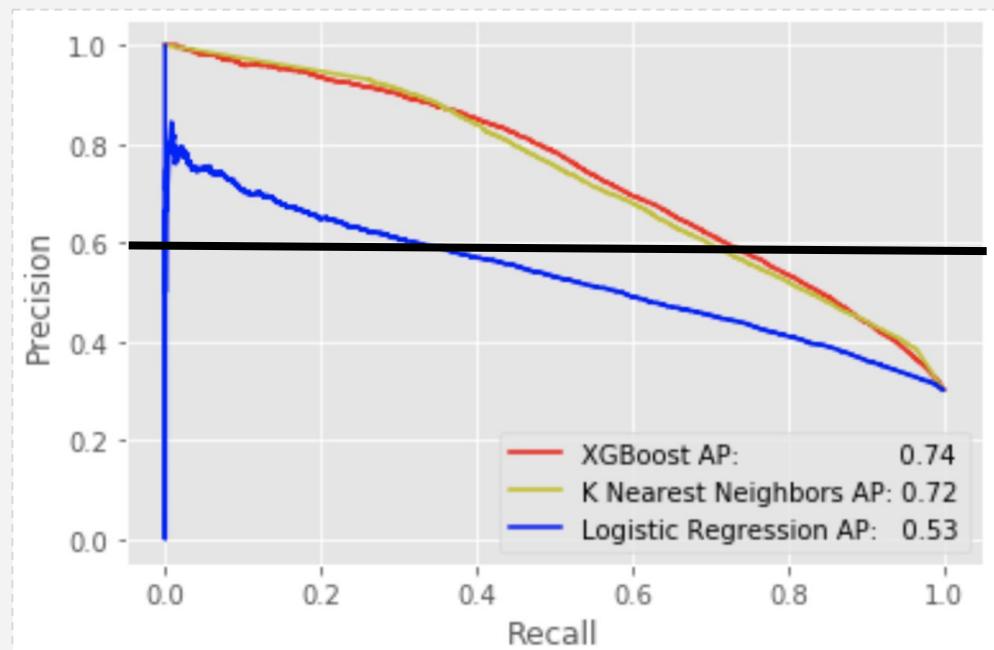


PRECISION RECALL CURVE



PRECISION RECALL CURVE

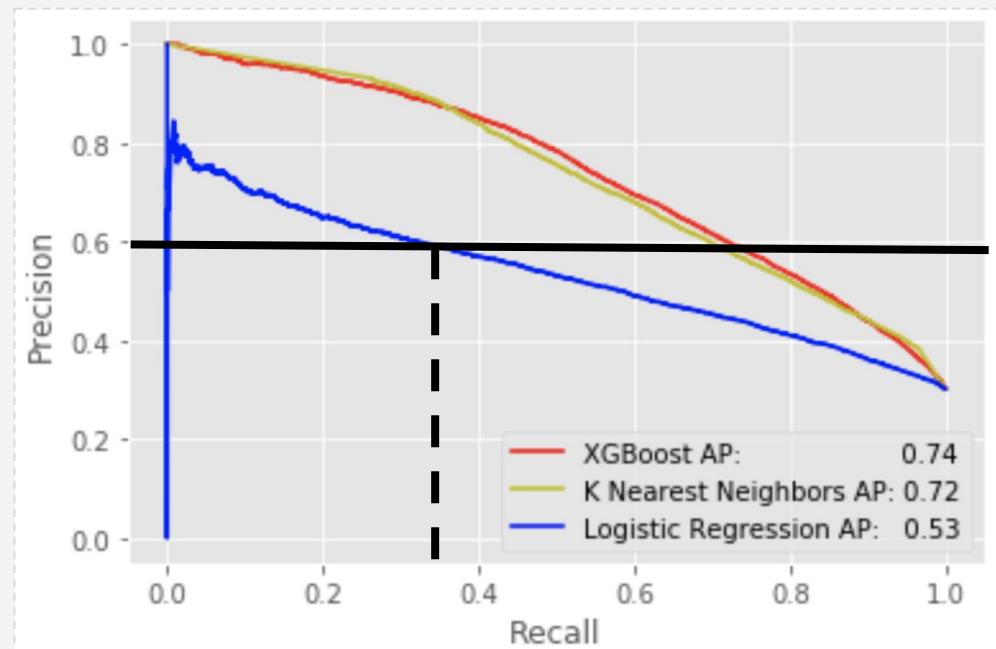
Choosing Precision 60%:



PRECISION RECALL CURVE

Choosing Precision 60%:

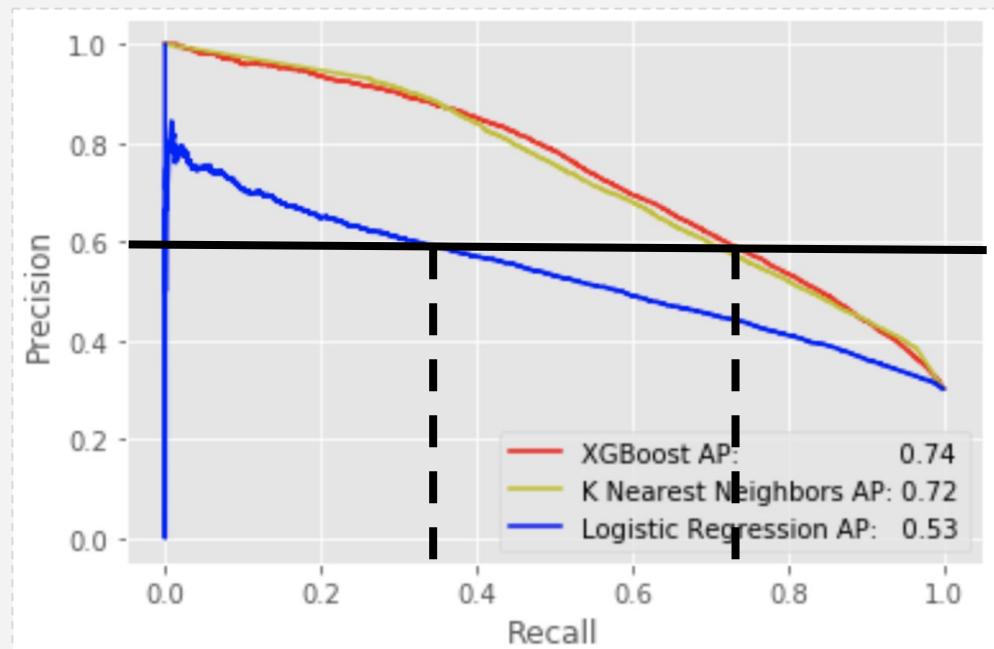
- Logistic Regression achieves **33% Recall**



PRECISION RECALL CURVE

Choosing Precision 60%:

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



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

