Reselling platform

Who will buy the second-hand luxury article in the consumer-to-consumer(C2C) platform?

Contents

1. Problem frame

    1-1. Problem

    1-2. Data set

    1-3. Exploratory data analysis

2. Model

    2-1. Baseline

    2-1. Improved model

3. Interpretation

    3-1 Interpretation of features

    3-2 Action proposal

1. Problem frame

1-1. Problem

Which user purchase an article in this platform?

Vestiaire Collective

AUTHENTICATED PRE-OWNED LUXURY FASHION

Vestiaire Collective is an online vintage mall.

For this C2C platform to succeed,
transaction volume should grow.

This time, let’s see **the characteristics of buyer**
in this platform.

Scraped user data from Vestiaire Collective

identifierHash
type
country
language
socialNbFollowers
socialNbFollows
socialProductsLiked
productsListed
productsSold
productsPassRate
productsWished
productsBought
gender
civilityGenderId
civilityTitle
hasAnyApp
hasAndroidApp
hasIosApp
hasProfilePicture
daysSinceLastLogin
seniority
seniorityAsMonths
seniorityAsYears
countryCode

98,913 users
24 features

identifierHash
type
country
language
socialNbFollowers
socialNbFollows
socialProductsLiked
productsListed
productsSold
productsPassRate
productsWished
productsBought
gender
civilityGenderId
civilityTitle
hasAnyApp
hasAndroidApp
hasIosApp
hasProfilePicture
daysSinceLastLogin
seniority
seniorityAsMonths
seniorityAsYears
countryCode

98,913 users
15 features

- Data set has been collected from Kaggle[1]
  - User data of Vestiaier Collective
- Before preprocessing
  - 98,913 users, 24 features
  - No missing values
  - No duplicate data
- Dropped 9 features
  - Redundant features: type, gender, civilityTitle, hasAnyApp, seniorityAsMonths, seniorityAsYears
  - Features of high cardinality: identifierHash, country, countryCode
- After preprocessing
  - **98,913 users, 15 features**

[1] E-commerce - Users of a French C2C fashion store (contributed by JEFFREY MVUTU MABILAMA)

1-2. Data set

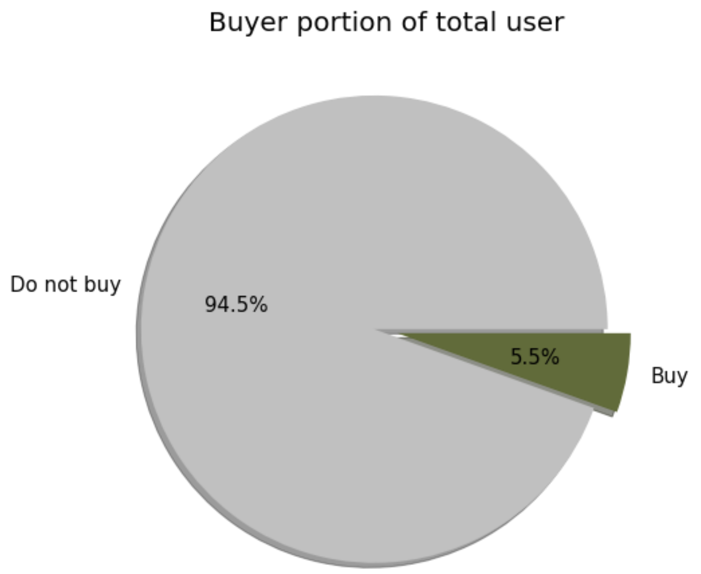
Scraped user data from Vestiaire Collective

Variable	Description[1]
language	The user's preferred language
socialNbFollowers	Number of users who follow this user's activity. New accounts are automatically followed by the store's official
socialNbFollows	Number of user account this user follows. New accounts are automatically assigned to follow the official partners
socialProductsLiked	Number of products this user liked
productsListed	Number of currently unsold products that this user has uploaded.
productsSold	Number of products this user has sold
productsPassRate	% of products meeting the product description. (Sold products are reviewed by the store's team before being shipped to the buyer)
productsWished	Number of products this user added to his/her wish list.
productsBought	**Number of products this user bought (Target of this analysis)**
civilityGenderId	1, 2, 3 (1 is Mr., 2 is Mrs, 3 is Miss)
hasAndroidApp	If user has ever used the official Android app
hasIosApp	If user has ever used the official iOS app
hasProfilePicture	If user has a custom profile picture
daysSinceLastLogin	Number of days since the last login
seniority	Number of days since the user registered

[1] EDA: Online C2C fashion store - user behaviour (Kaggle, JEFFREY MVUTU MABILAMA)

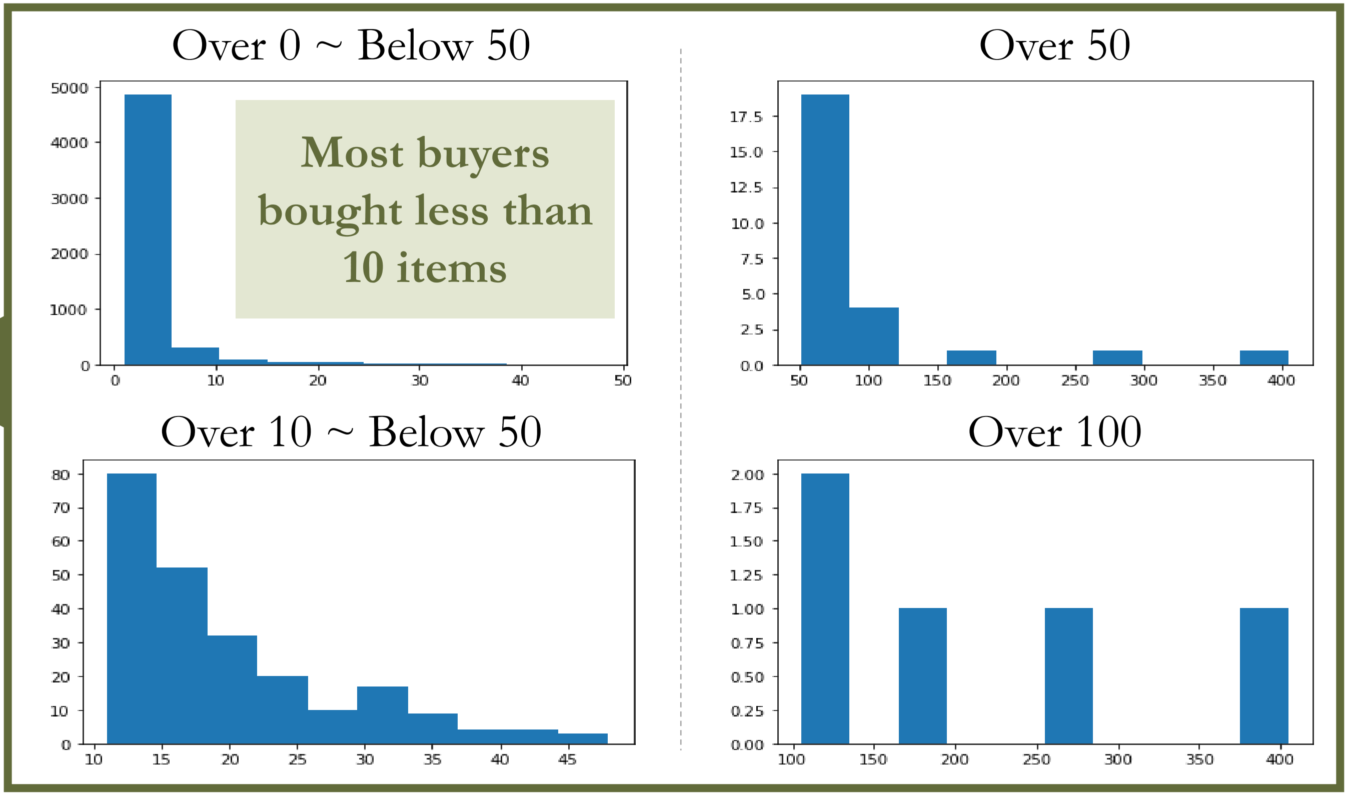
Target is imbalanced and right skewed

5.5% of total users ever bought an item



Total	98,913
Do not buy	93,494
Buy	5,419

Distribution of buyers by the number of purchased item



Over 0 ~ Below 50

Over 50

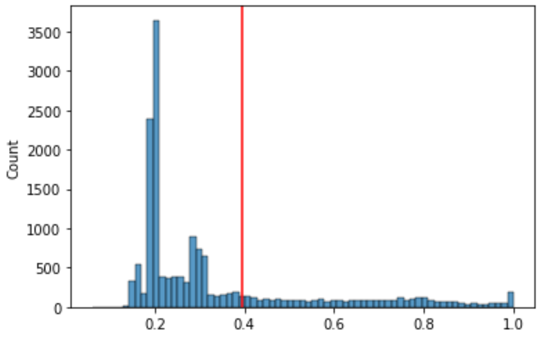
Over 10 ~ Below 50

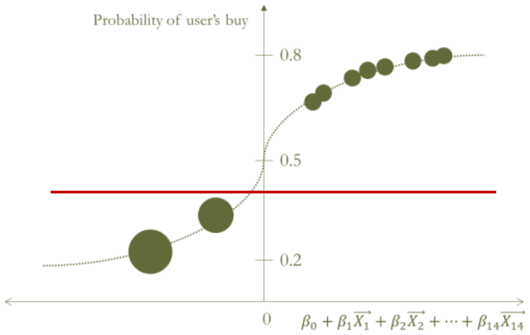
Over 100

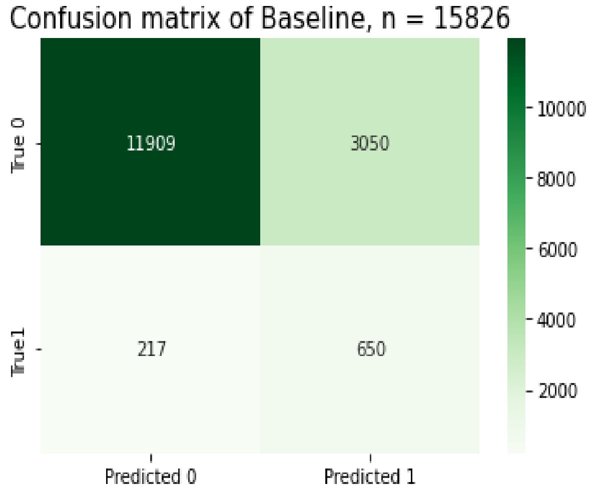
Most buyers bought less than 10 items

2. Model

Logistic regression, recall is 0.75




Probability of user's buy
0.8
0.5
0.2
0
$\beta_0 + \beta_1\overrightarrow{X_1} + \beta_2\overrightarrow{X_2} + \cdots + \beta_{14}\overrightarrow{X_{14}}$



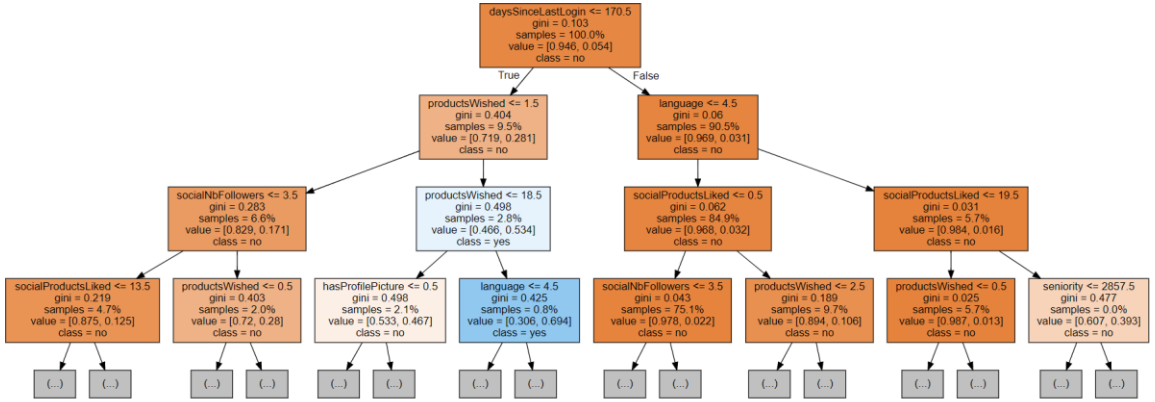
- Logistic regression is a model that determines probabilities by putting a function called sigmoid on the linearity of a feature and a target.
- Binary classification based on the threshold
- Optimal threshold [1] is calculated as 0.39, and users with a purchase probability of 39% or more are classified as buyers.

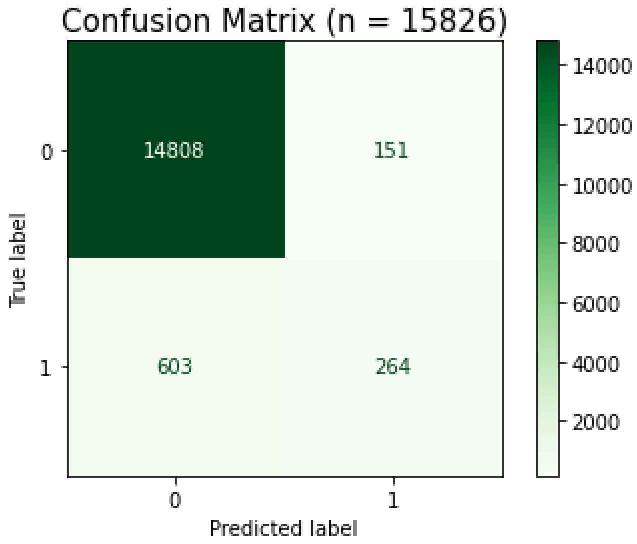
(Recall is 0.75)
This model retrieved **75%** of buyers.
(Fail in retrieving 25% of buyers.)

[1] Use Area Under Curve (AUC) as the optimal threshold criterion

Random Forest, recall is 0.3

One of many trees



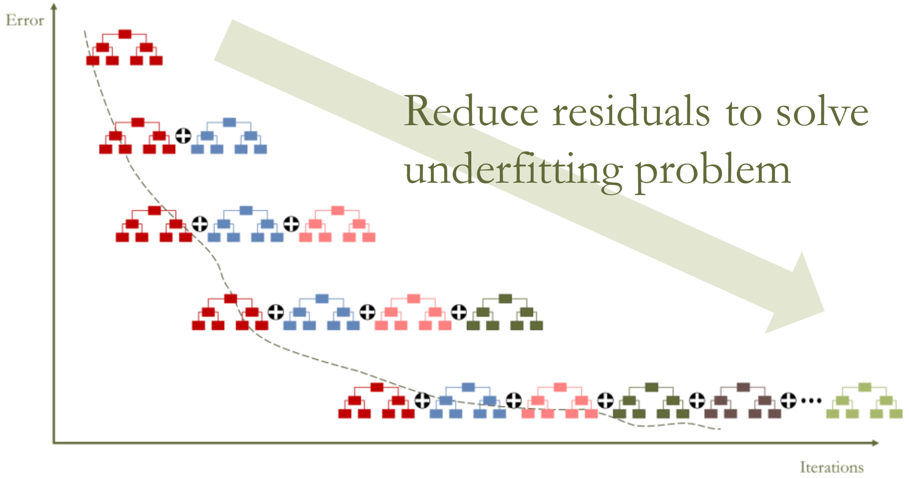


- Random Forest creates multiple trees and classifies users as purchasers/non-buyers by majority vote.
- Sampling with restoration was used, and the features are randomly extracted to create a tree, thereby alleviating the overfitting problem of trees that fit only a specific data set.

(Recall is 0.3)
This model retrieved 30% of buyers.
(Fail in retrieving 70% of buyers.)

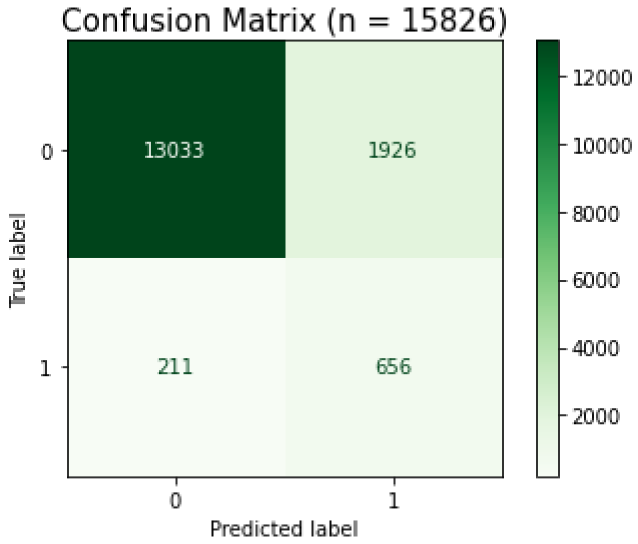
2-2. Improved model

Gradient boosting decision tree, recall is 0.76



Reduce residuals to solve underfitting problem

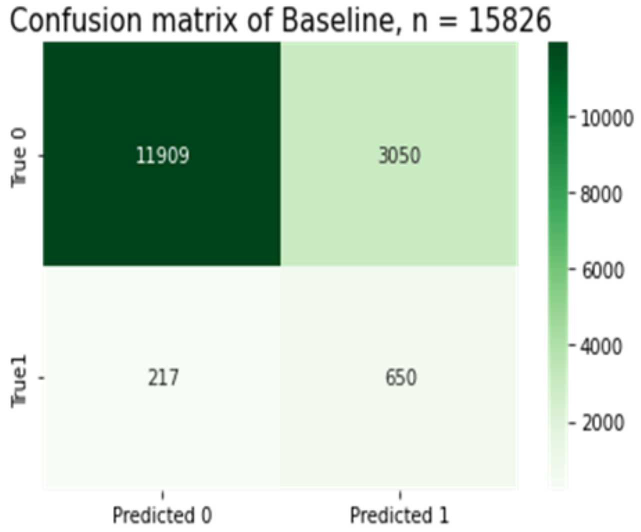
- Gradient boosting decision tree is also an ensemble model that generates multiple trees to predict targets.
- Alleviate the overfitting problem by limiting the number of leaves in the tree
- Also alleviate the underfitting problem by continuing to create a tree that reduces the residual of the tree created previously.



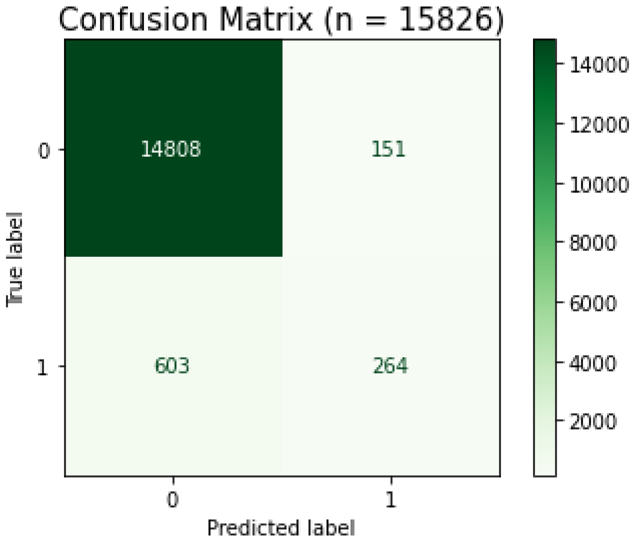
(Recall is 0.76)
This model retrieved **76%** of buyers.
(Fail in retrieving 24% of buyers.)

2-2. Improved model

Performance of the gradient boosting decision tree model has slightly improved

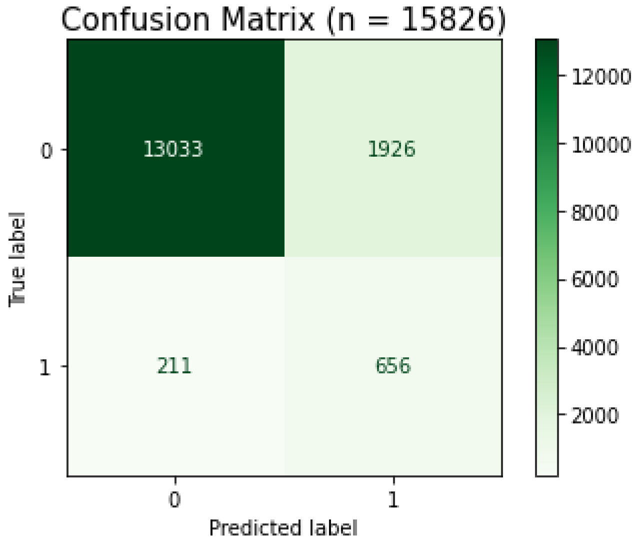


Logistic regression (baseline) recall **0.75**



Random forest recall **0.3**

Final model



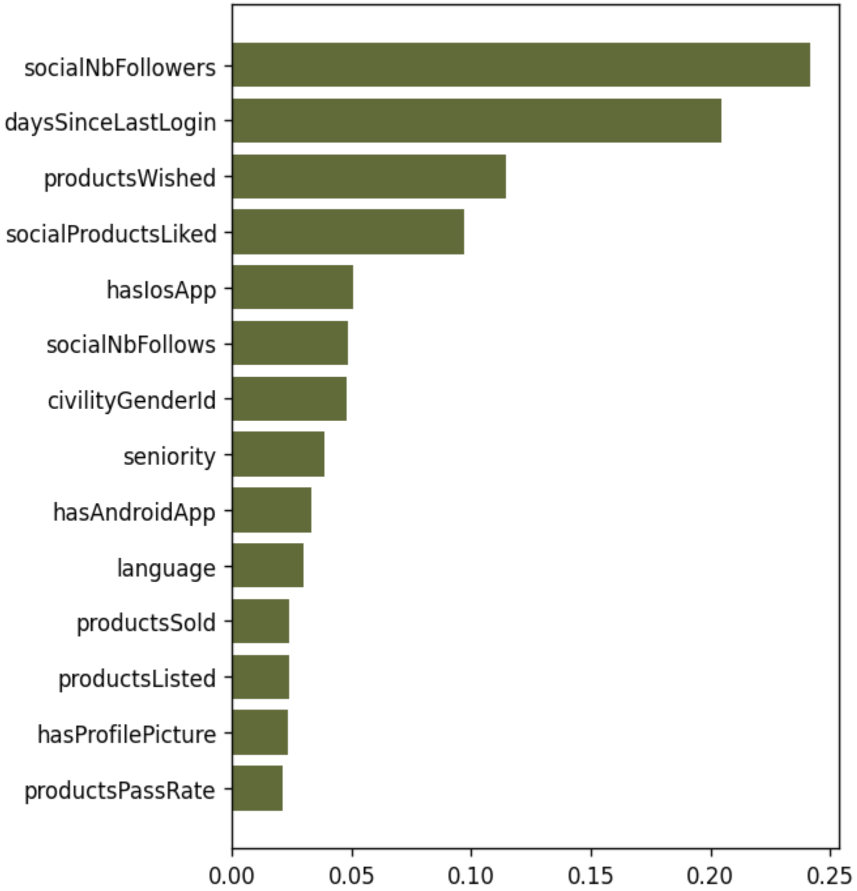
Gradient boosting decision tree recall **0.76**

3. Interpretation

4 characteristics that are highly relevant to the probability of purchase

More relevant feature

Less relevant feature



Among feature importance [1], the difference between 'hasIosApp' and 'socialProductsLiked' is drastic as twicse.

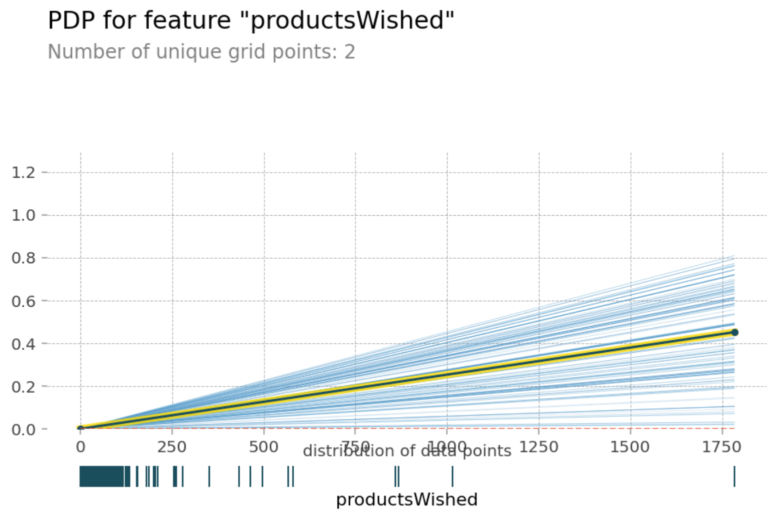
Therefore, this time, the features that are highly related to the probability of purchase are determined as the following four points.

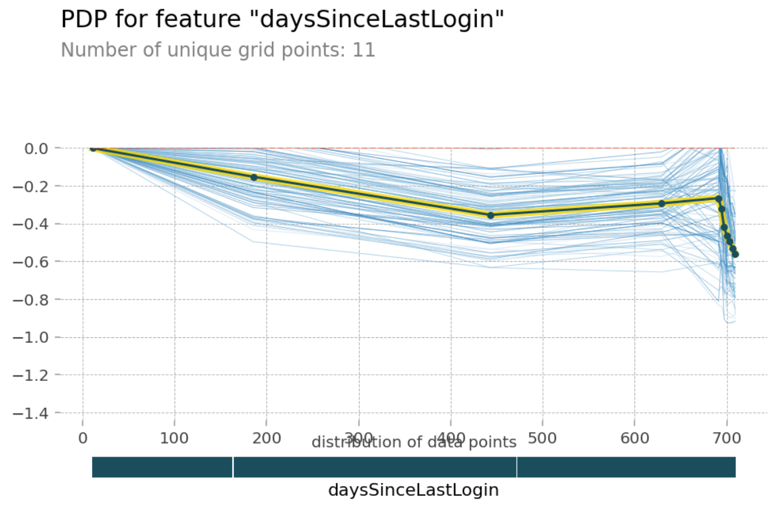
- ✓ Number of users who follow this user's activity.
- ✓ Number of days since the last login.
- ✓ Number of products this user added to his/her wish list.
- ✓ Number of products this user liked.

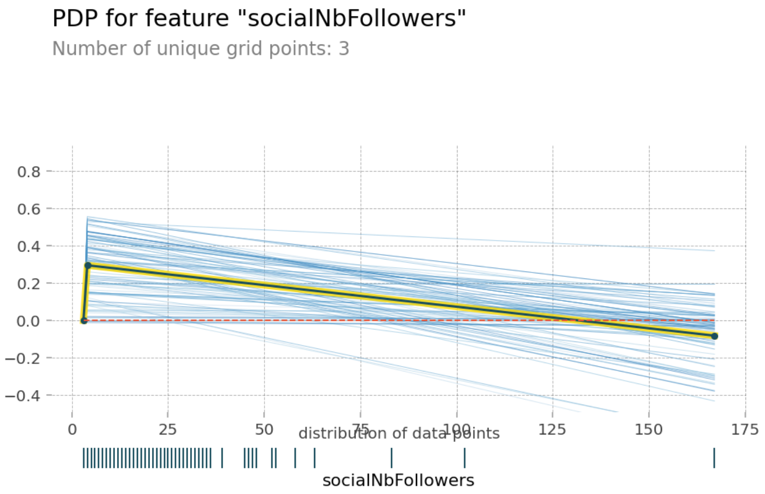
Note: Positive or negative correlation is unknown, and causation is unknown

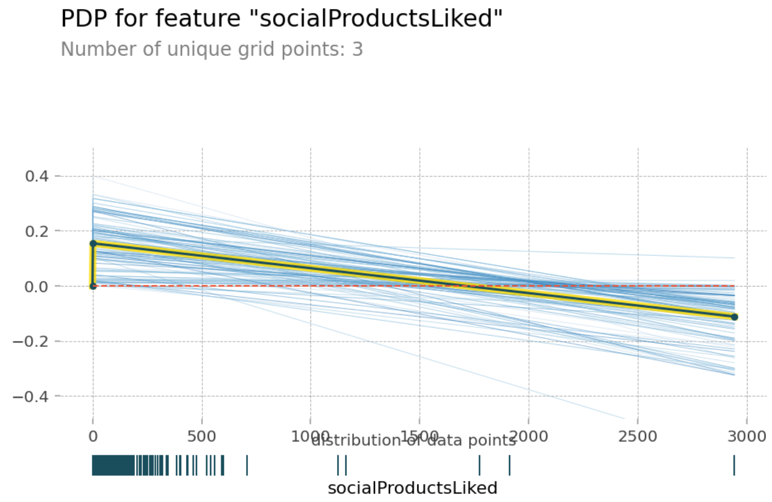
[1] Feature importance is a criterion for checking how much the performance of the model changes when certain features are lost or differ from the original data, to determine how importantly the features are related.

Partial dependence plot (PDP) and random 100 individual conditional expectation curves



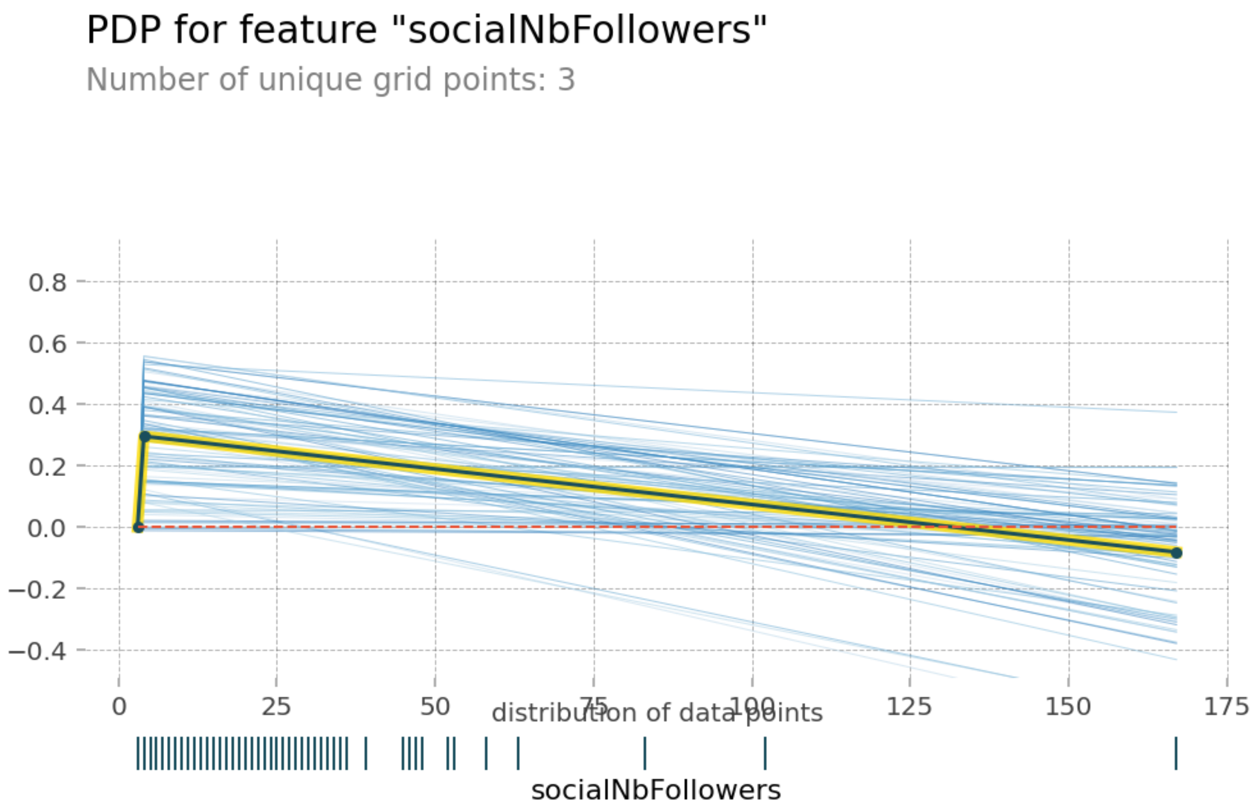




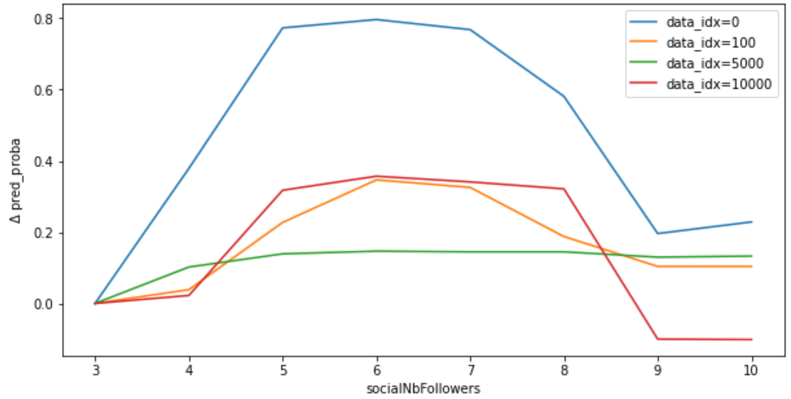


3-1. Feature interpretation

Relationship between *counts of social follower* and *purchase probability*

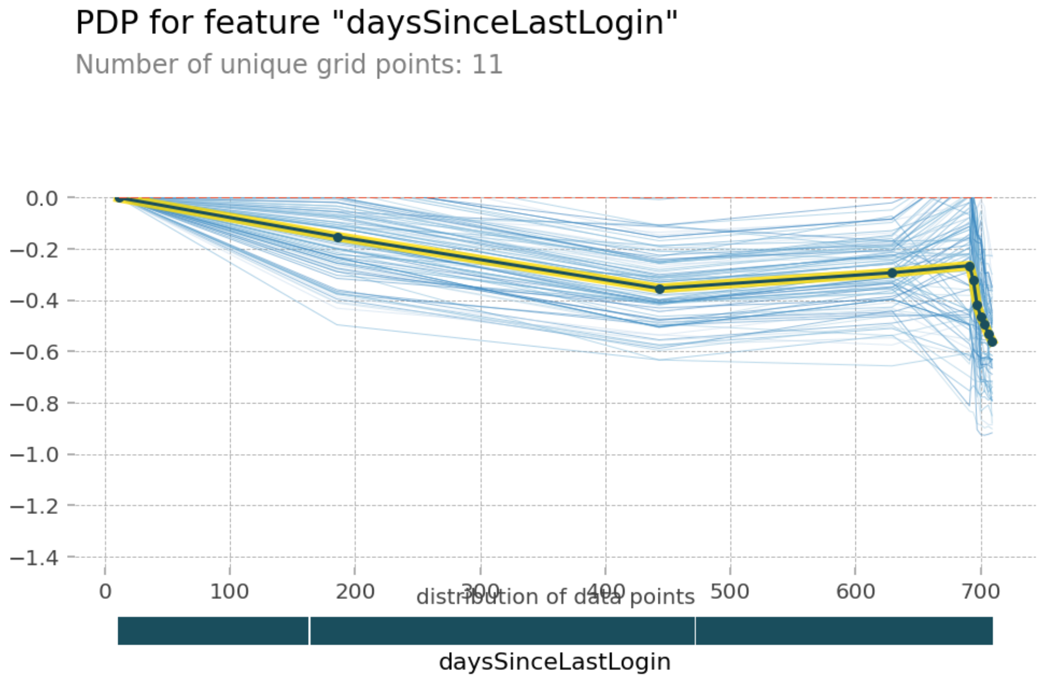


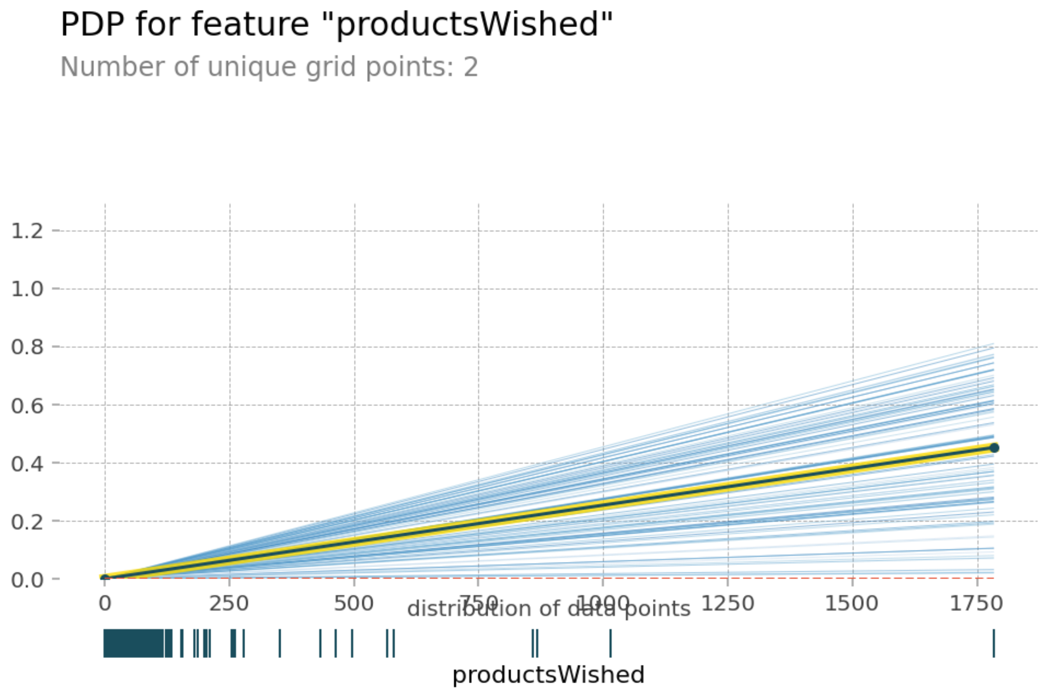
- Only with 1~2 increase of social followers, the probability of being a buyer increases
- However, with the more of social followers, the probability decreases



- Causal relationship between social follower and the purchase needs to be further investigated.
- The current hypothesis is that sellers often follow buyers. Most buyers buy less than 5 pieces, so if the seller follows, the number makes sense.

Reasonable relationship between both *days past since log-in* and *wished product* and *purchase probability*

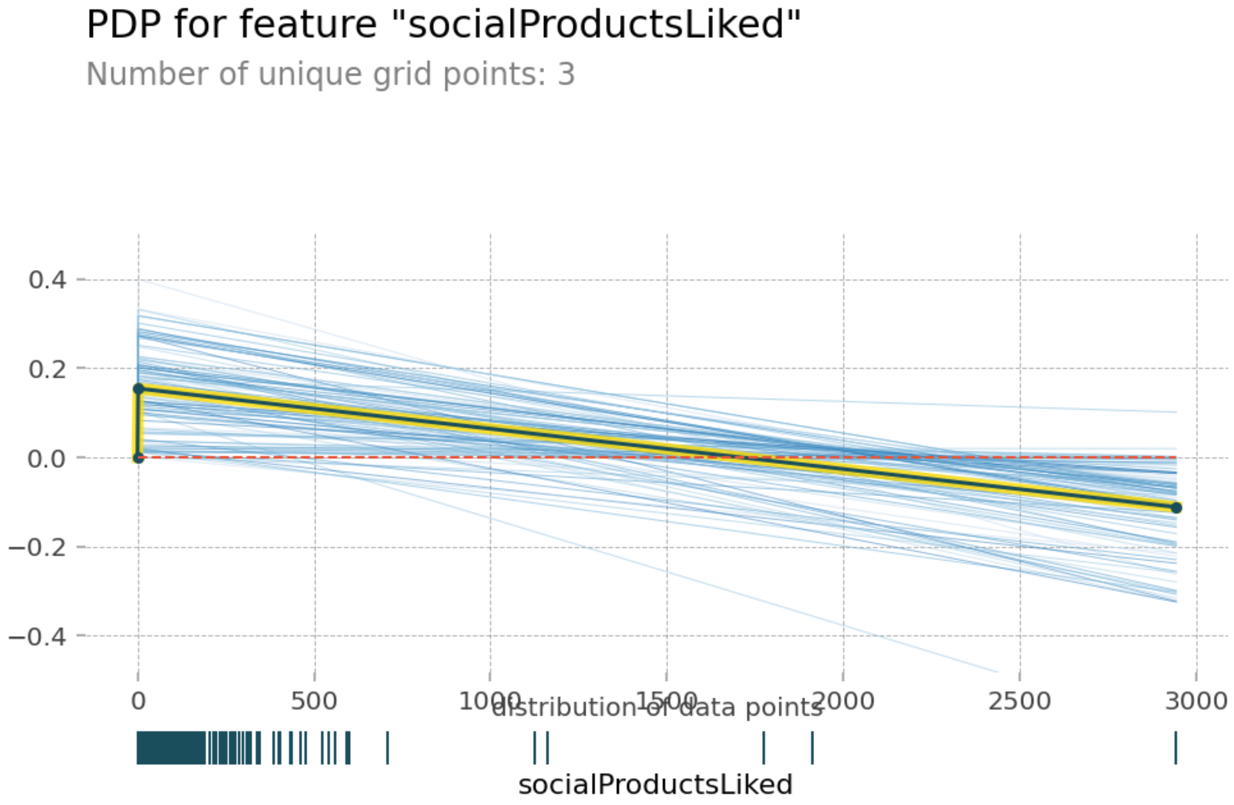




It is common sense that active users who have recently logged in are more likely to make purchases, and users with a lot of interested products are more likely to make purchases.

3-1. Feature interpretation

Relationship between ***the number of products that users 'liked*** and ***the probability of purchase***


PDP for feature "socialProductsLiked"

Number of unique grid points: 3

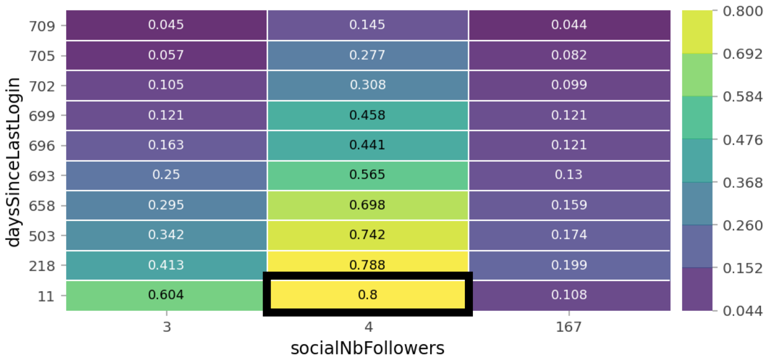
- It is contrary to common sense that the more ‘likes’ you click, the lower the probability of making a purchase.
- It is impossible to get too many likes, which seems to be the result of ICE making incorrect predictions.
- Further research is required to make an accurate judgment

3-1. Feature interpretation

Cases when the probability of purchase is greater than 80% in PDPs of both characteristics

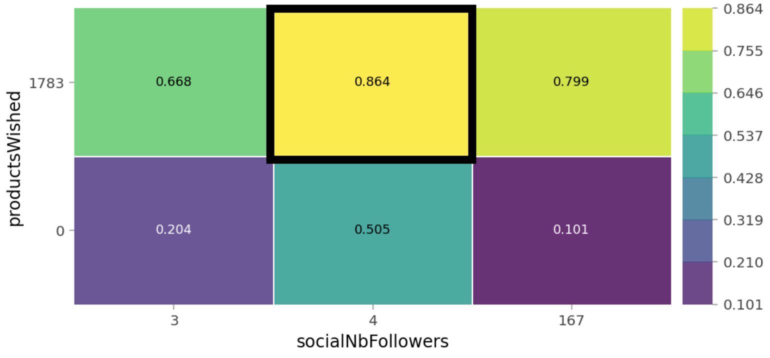
PDP interact for "socialNbFollowers" and "daysSinceLastLogin"

Number of unique grid points: (socialNbFollowers: 3, daysSinceLastLogin: 10)



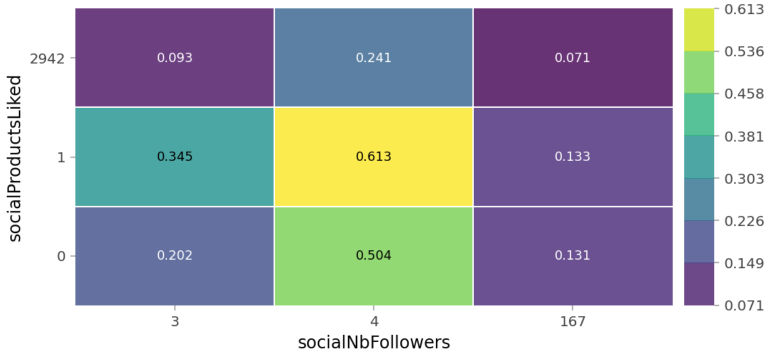
PDP interact for "socialNbFollowers" and "productsWished"

Number of unique grid points: (socialNbFollowers: 3, productsWished: 2)



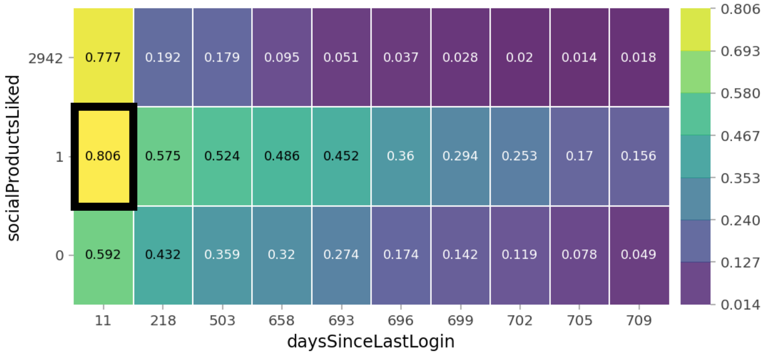
PDP interact for "socialNbFollowers" and "socialProductsLiked"

Number of unique grid points: (socialNbFollowers: 3, socialProductsLiked: 3)



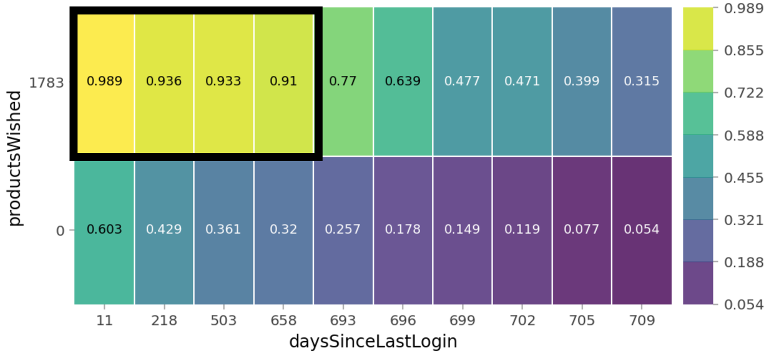
PDP interact for "daysSinceLastLogin" and "socialProductsLiked"

Number of unique grid points: (daysSinceLastLogin: 10, socialProductsLiked: 3)



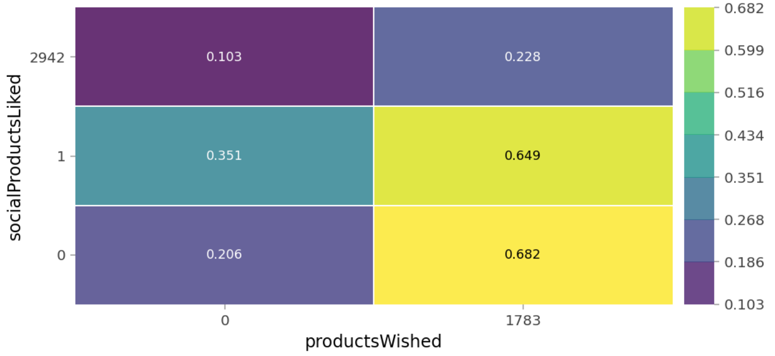
PDP interact for "daysSinceLastLogin" and "productsWished"

Number of unique grid points: (daysSinceLastLogin: 10, productsWished: 2)



PDP interact for "productsWished" and "socialProductsLiked"

Number of unique grid points: (productsWished: 2, socialProductsLiked: 3)

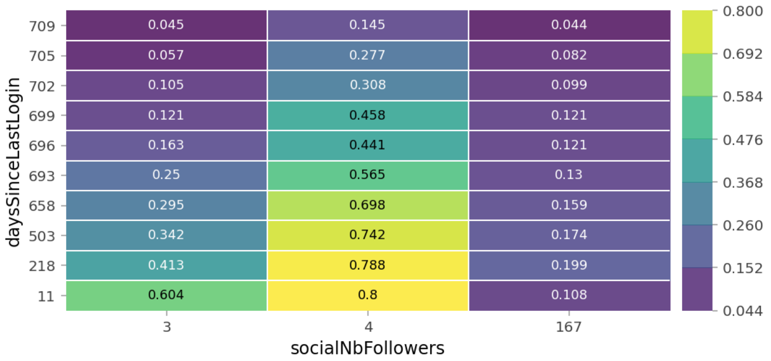


3-1. Feature interpretation

The more products customer wishes, the more likely customer is to buy, so you need to increase customer engagement

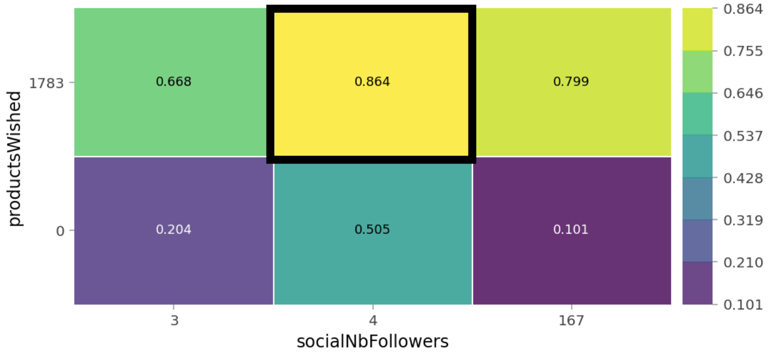
PDP interact for "socialNbFollowers" and "daysSinceLastLogin"

Number of unique grid points: (socialNbFollowers: 3, daysSinceLastLogin: 10)



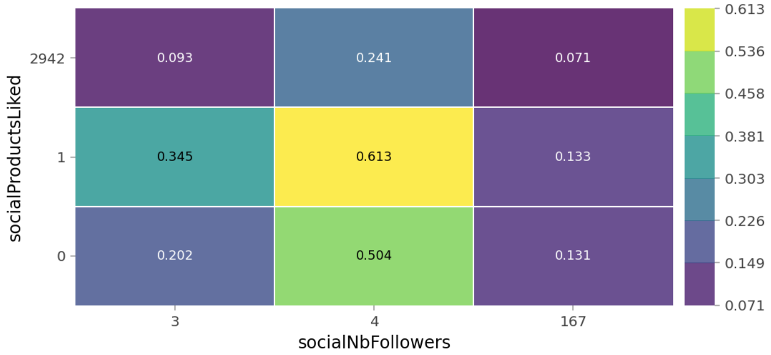
PDP interact for "socialNbFollowers" and "productsWished"

Number of unique grid points: (socialNbFollowers: 3, productsWished: 2)



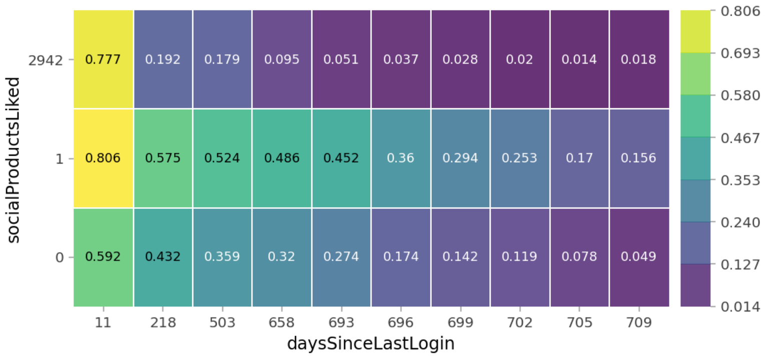
PDP interact for "socialNbFollowers" and "socialProductsLiked"

Number of unique grid points: (socialNbFollowers: 3, socialProductsLiked: 3)



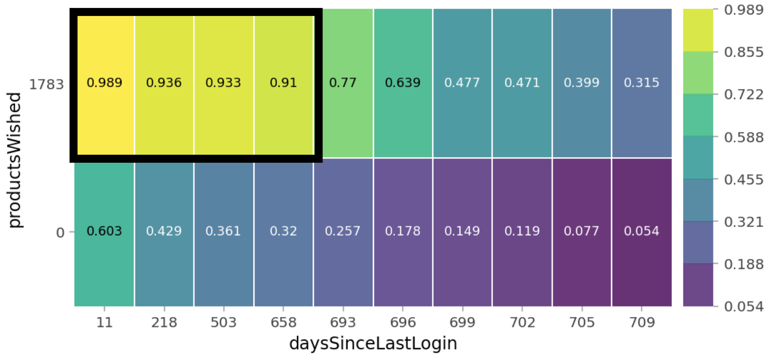
PDP interact for "daysSinceLastLogin" and "socialProductsLiked"

Number of unique grid points: (daysSinceLastLogin: 10, socialProductsLiked: 3)



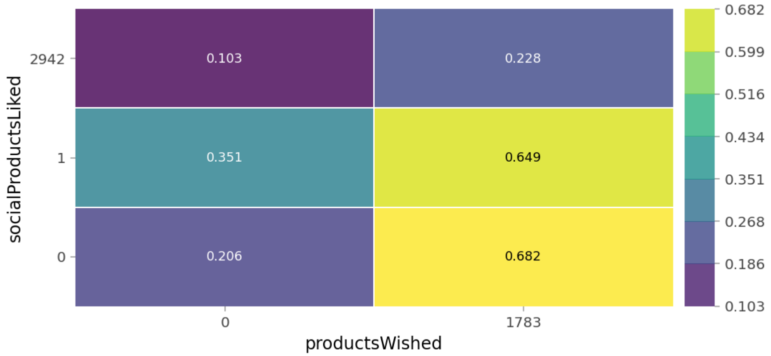
PDP interact for "daysSinceLastLogin" and "productsWished"

Number of unique grid points: (daysSinceLastLogin: 10, productsWished: 2)



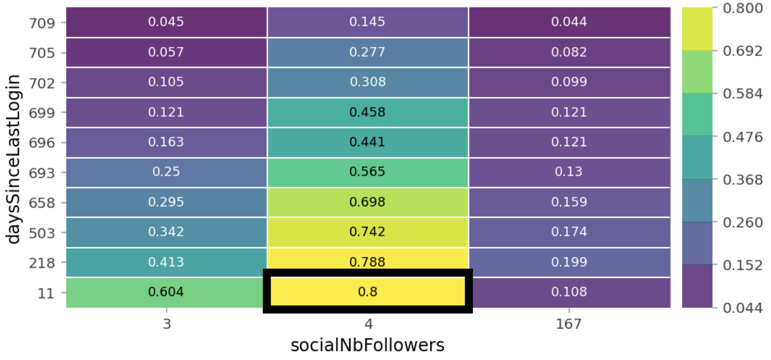
PDP interact for "productsWished" and "socialProductsLiked"

Number of unique grid points: (productsWished: 2, socialProductsLiked: 3)

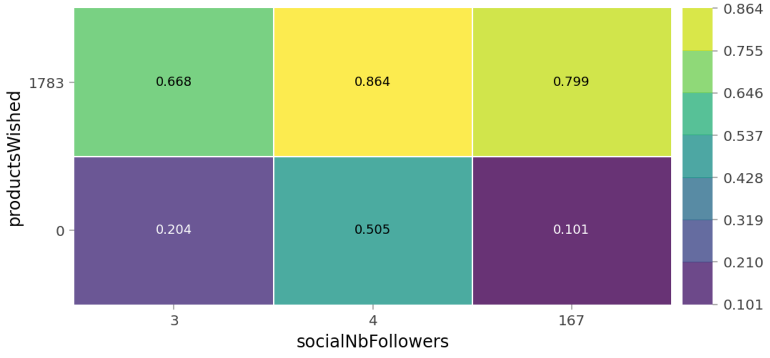


3-1. Feature interpretation

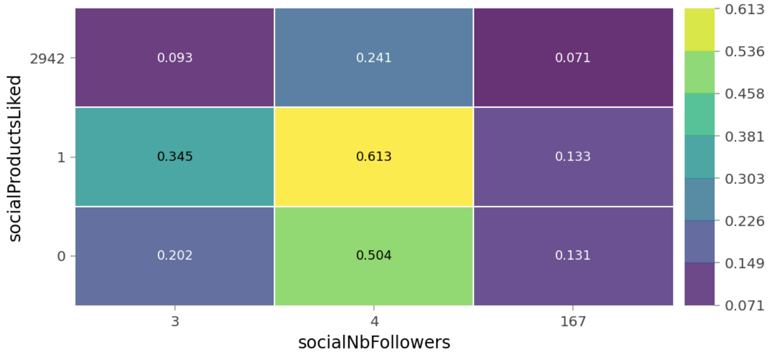
The more recently a customer logs in, the higher the probability of purchases is, so you need to increase the activity of your account.


PDP interact for "socialNbFollowers" and "daysSinceLastLogin"

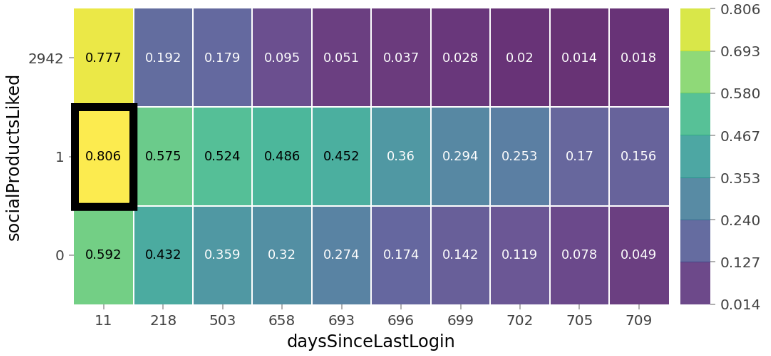
Number of unique grid points: (socialNbFollowers: 3, daysSinceLastLogin: 10)


PDP interact for "socialNbFollowers" and "productsWished"

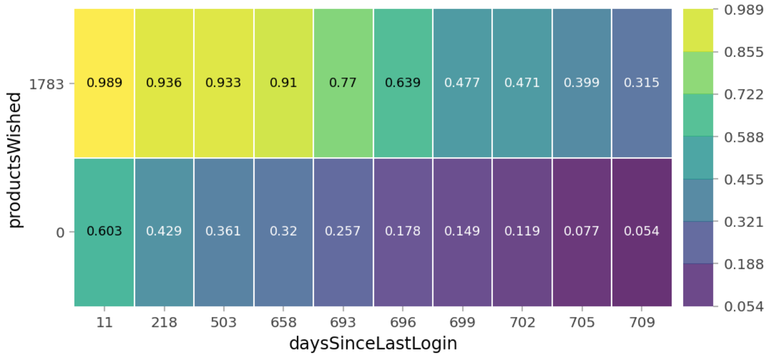
Number of unique grid points: (socialNbFollowers: 3, productsWished: 2)


PDP interact for "socialNbFollowers" and "socialProductsLiked"

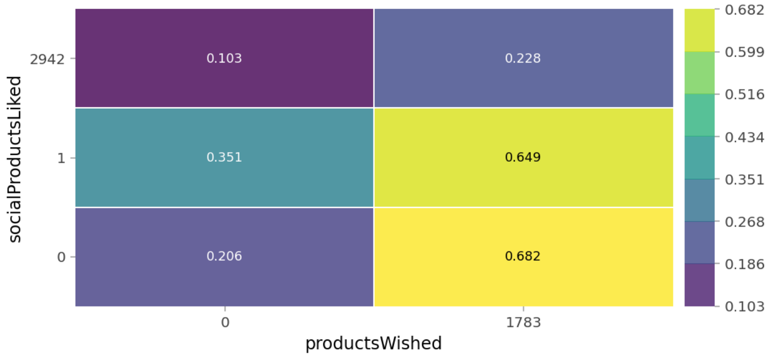
Number of unique grid points: (socialNbFollowers: 3, socialProductsLiked: 3)


PDP interact for "daysSinceLastLogin" and "socialProductsLiked"

Number of unique grid points: (daysSinceLastLogin: 10, socialProductsLiked: 3)


PDP interact for "daysSinceLastLogin" and "productsWished"

Number of unique grid points: (daysSinceLastLogin: 10, productsWished: 2)


PDP interact for "productsWished" and "socialProductsLiked"

Number of unique grid points: (productsWished: 2, socialProductsLiked: 3)

3 suggestions

1. Focus on efforts to bring customers to the intent phase in Marketing funnel.
2. Makes customers want to log in constantly in variable ways such as providing interesting content.
3. Since the probability of purchase does not increase with more social product likes, we should redesign the like so that it can be an indicator of purchase intent.