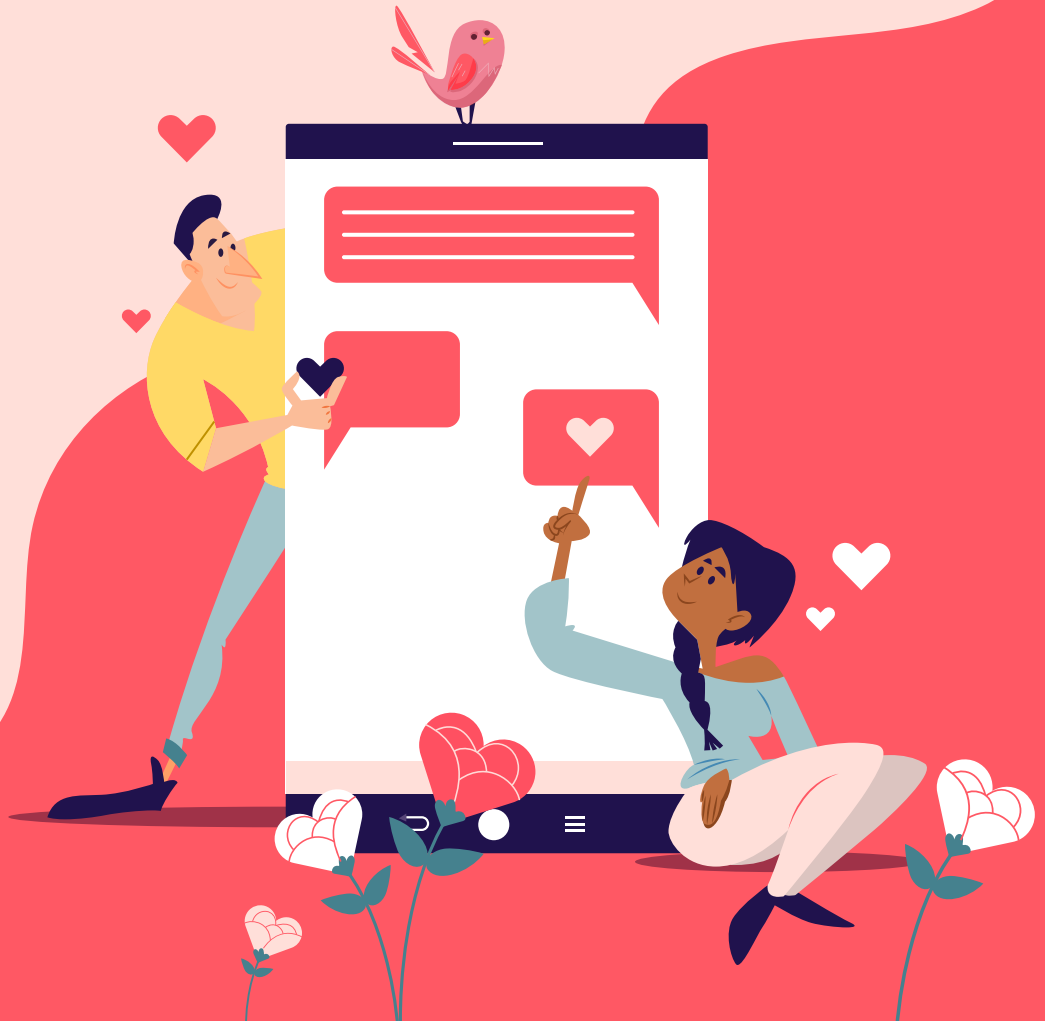
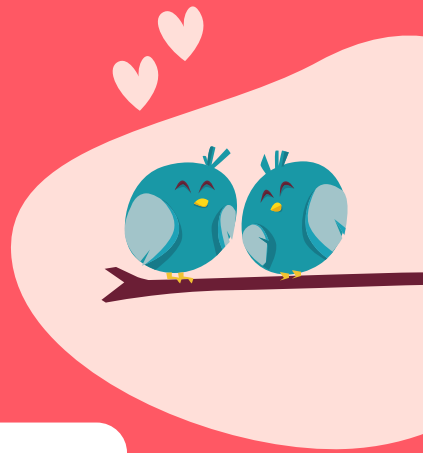
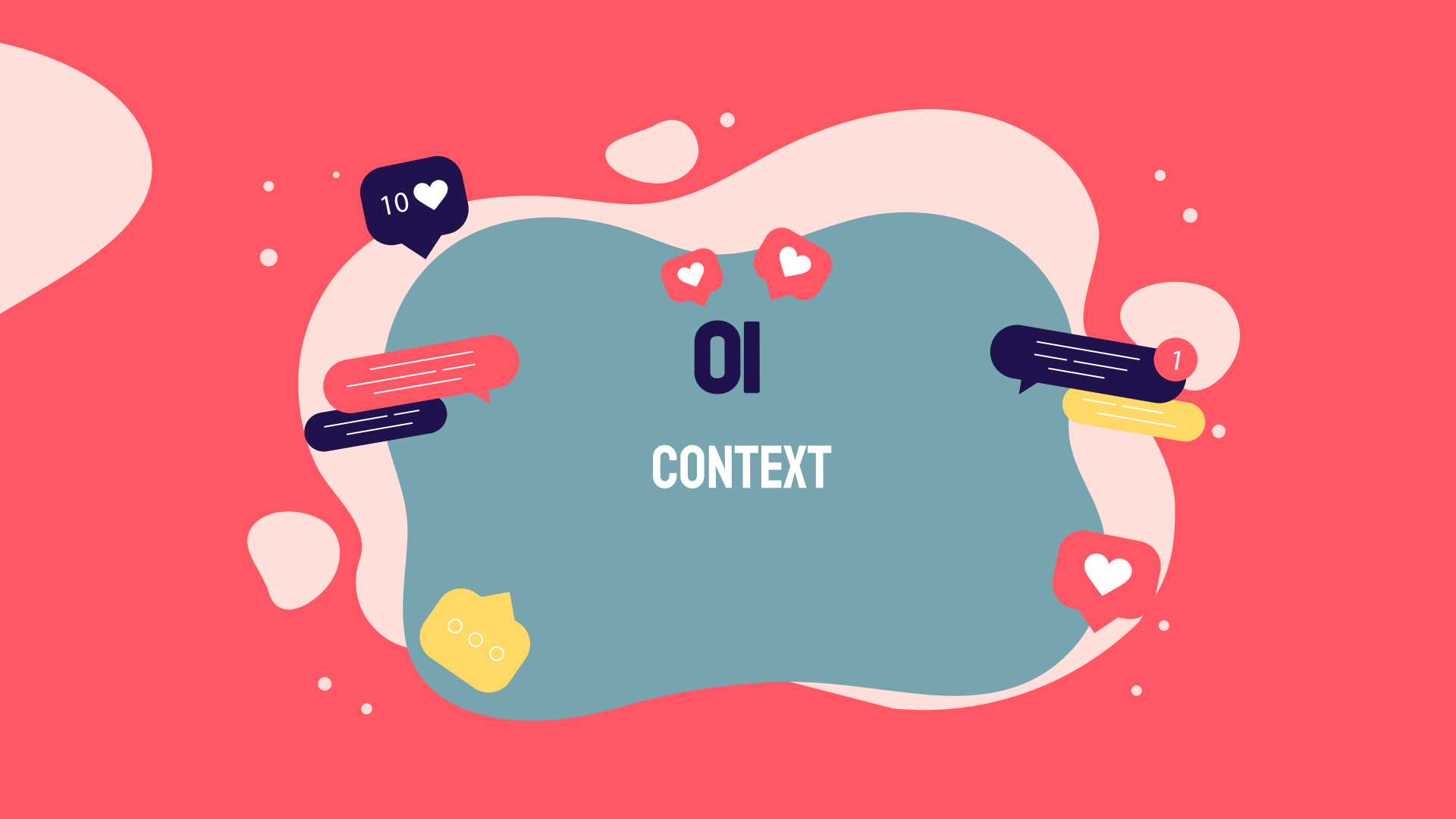
# CHURN PREDICTION FOR A DATING APPLICATION



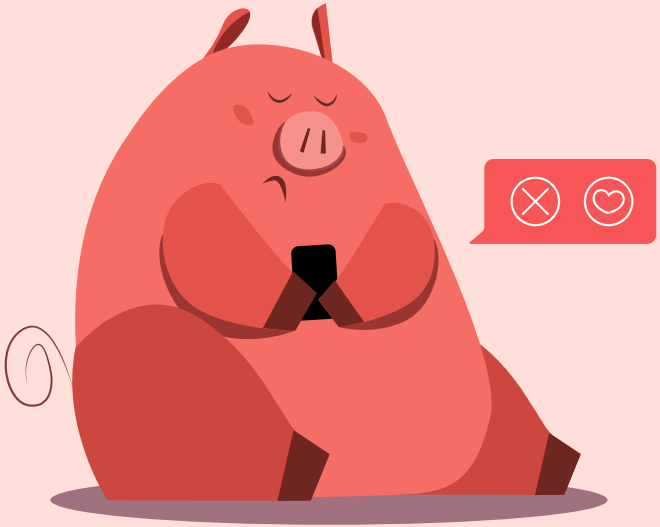
# TABLE OF CONTENTS




10
01
CONTEXT
1

# APPLICATION DESCRIPTION

**For confidential reason, this part of the presentation has been removed**



# PROBLEM

**“An increase by 5% of customer retention can increase profit by 25% to 95 %”**

*Bain & company study (Frederick F. Reichheld and Phil Schefter)*

**=> Creation of a model to predict churn of the application’s users**


10
02
PROCESS
1

# PROCESS


**AWS REDSHIFT**

**BI PLATFORM**

SQL Queries to csv files

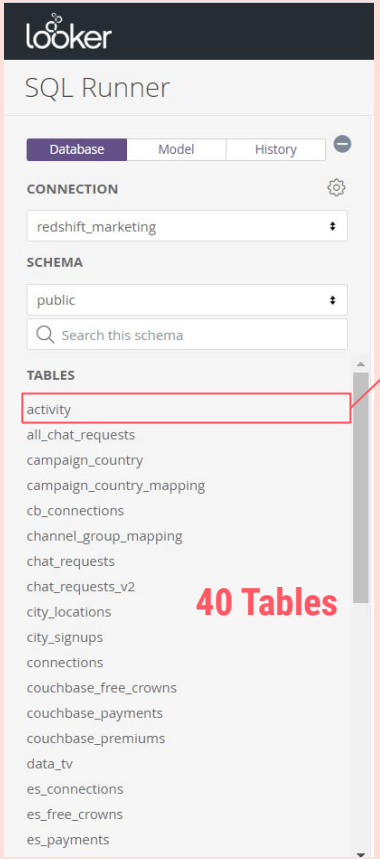
**DATA CLEANING**

**EXPLORATION**

**DATA PROCESSING**

**MODELING**

# DATA COLLECTION




| event_type | event_time | user_id | user_location_lat | user_location_lon |
|---|---|---|---|---|
| activity | 2017-09-28T15:21:38.000+00:00 | EA2618719 | 41.1171432 | 16.8718715 |
| activity | 2017-09-28T15:21:52.000+00:00 | EA4278277 | 48.2791461 | 10.9722113 |
| activity | 2017-09-28T15:21:52.000+00:00 | EA4895146 | 49.4911632 | 0.1155702 |
| activity | 2017-09-28T15:21:54.000+00:00 | EA4980827 | 38.7450669 | -77.6964566 |

Heavy SQL query to get the target : churn

Many SQL queries to get all necessary data

Queries pushed to csv => import to Python

# DATA COLLECTED

**LAST 2 MONTHS NEW USERS : 120K**

## USER PERSONAL INFORMATION

- ★ Gender
- ★ Age
- ★ Country
- ★ Platform

## USER BEHAVIOUR IN APP

- ★ Logins
- ★ Connections
- ★ Chat request sent / received
- ★ Purchases
- ★ Virtual currency spending
- ★ Number of ratings given
- ★ Average rating given

## OTHER

- ★ Acquisition mean (paid/organic)
- ★ Number of ratings received
- ★ Average rating received

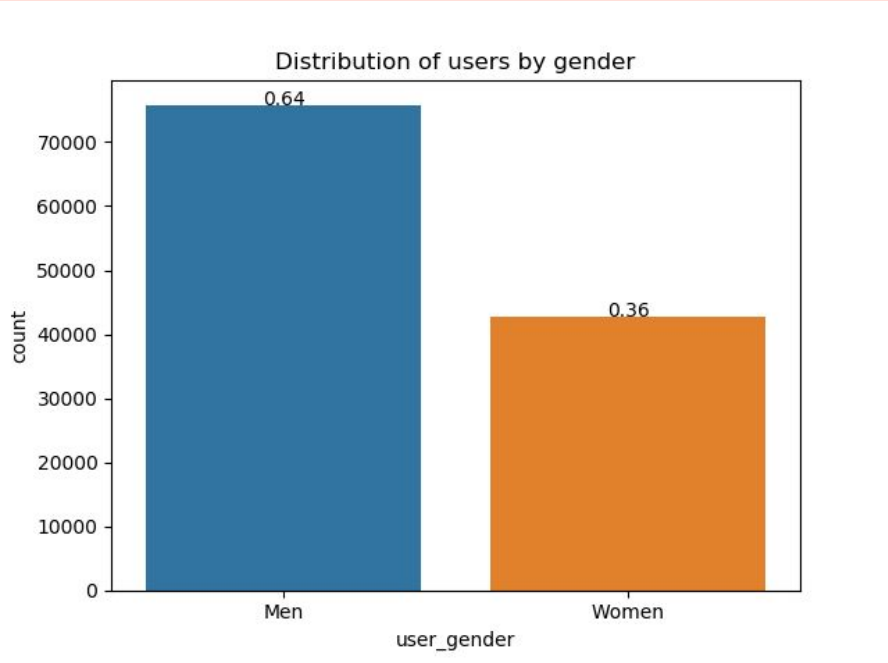
**BE CAREFUL WITH THE SCAMS !**

# FINAL OBJECTIVE

Predict the user that are not going to show up in week 2 according to their inapp experience during the 5 first days.

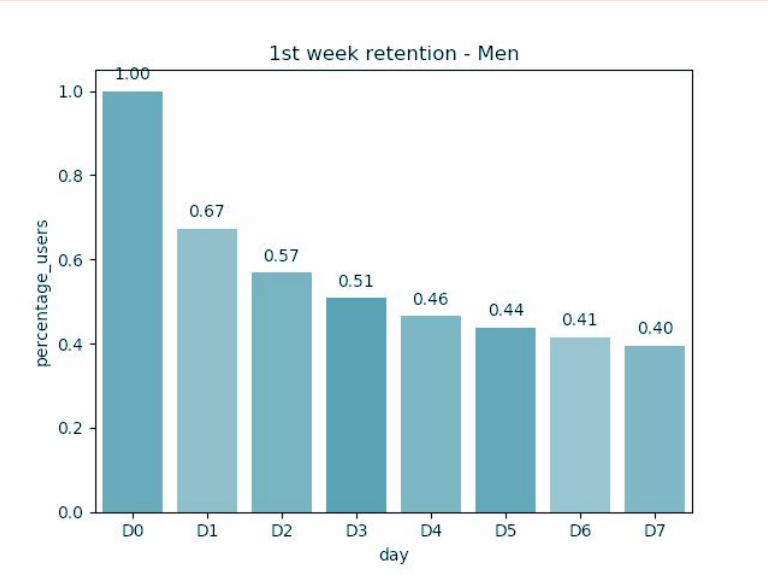

10
03
EXPLORATION
1

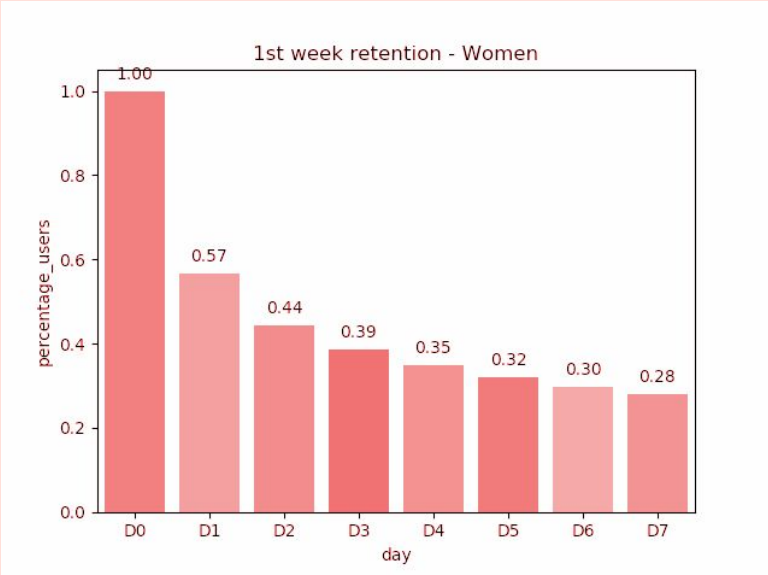
# GENDER REPARTITION




The gender repartition is imbalanced

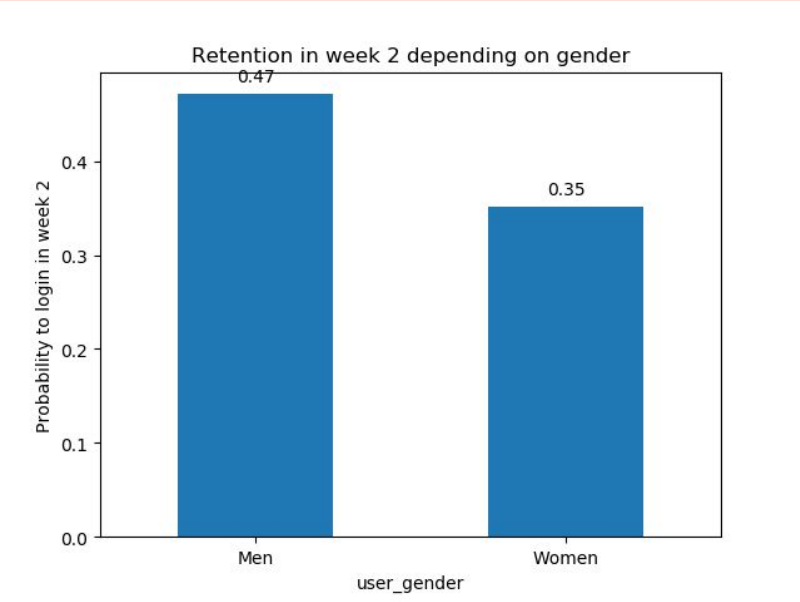
# FIRST WEEK RETENTION



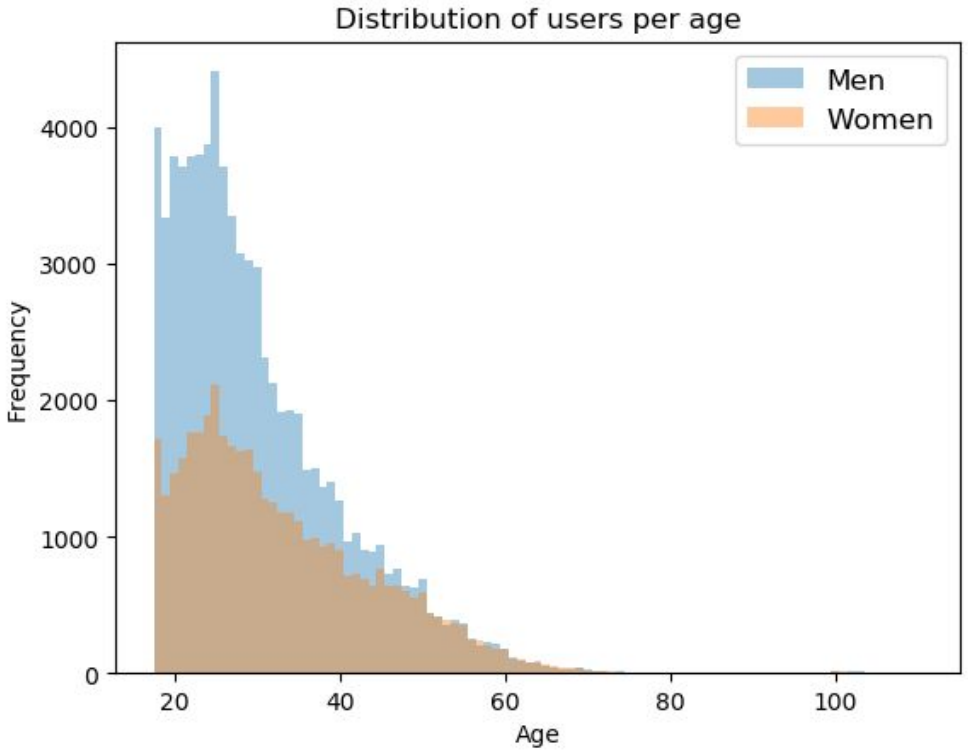




Women drop the app more easily than men.
And since they are fewer at the beginning it leads to a even stronger imbalanced repartition of gender !

# CHURN PROBABILITY BY GENDER

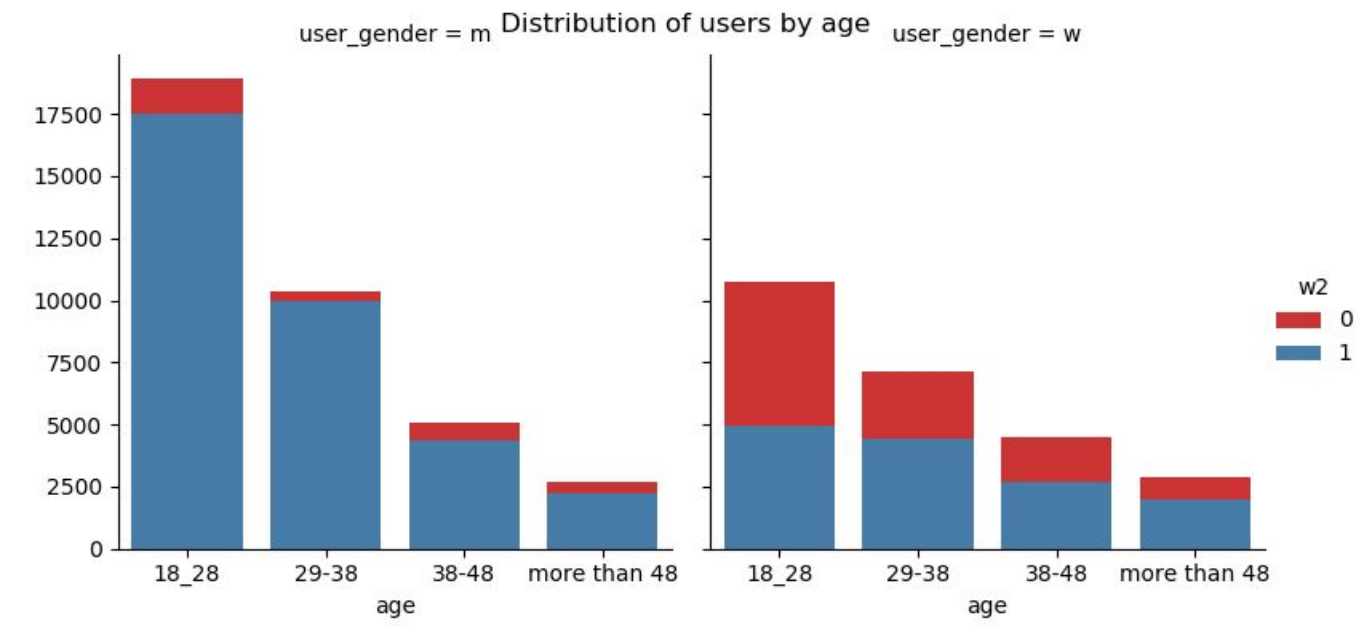


# USER AGE DISTRIBUTION



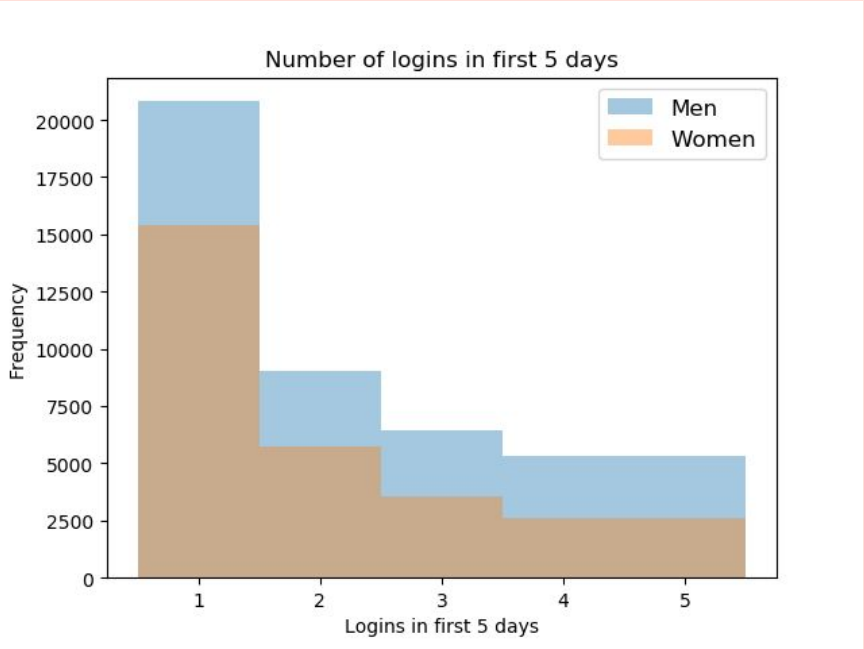

The majority of the users are in their late 20's - early 30's.

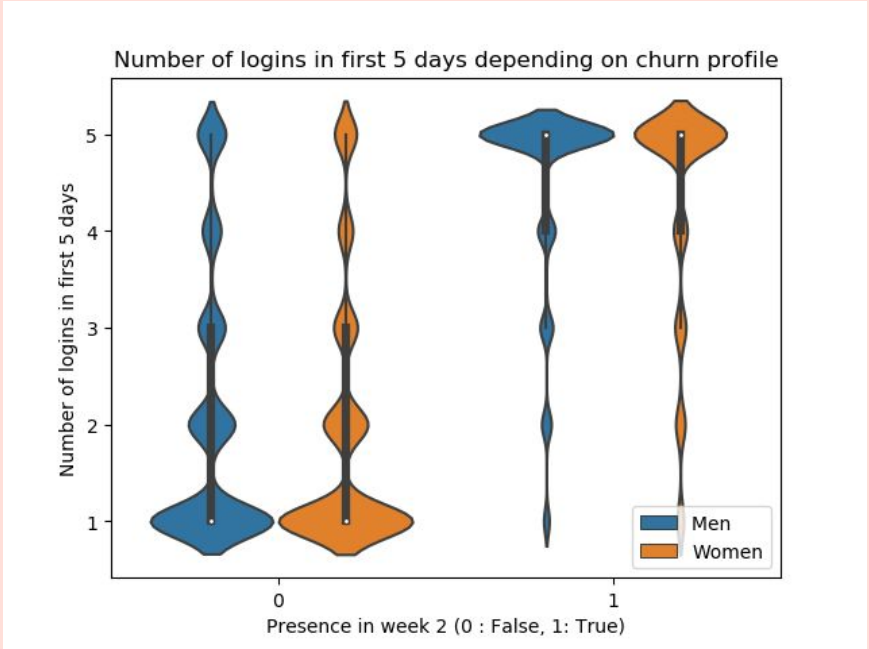
# IMPACT OF AGE ON CHURN




Younger users tend to churn more easily, especially the women !

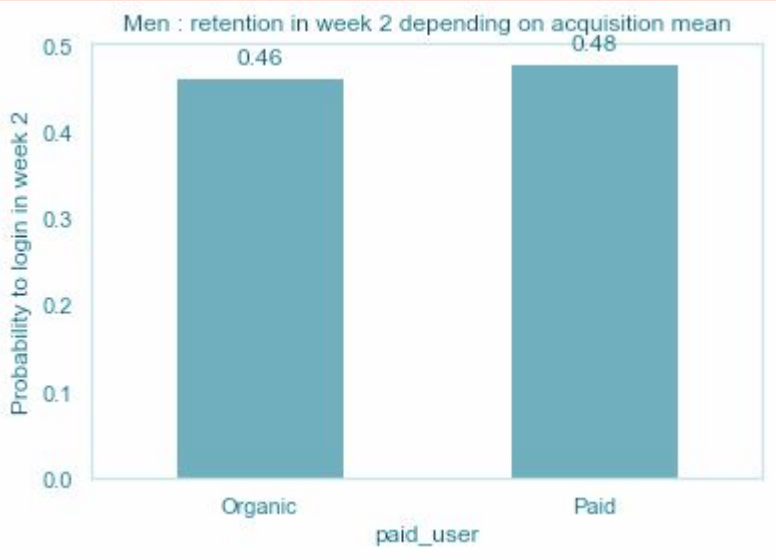
# IMPACT OF EARLY FREQUENCY ON CHURN

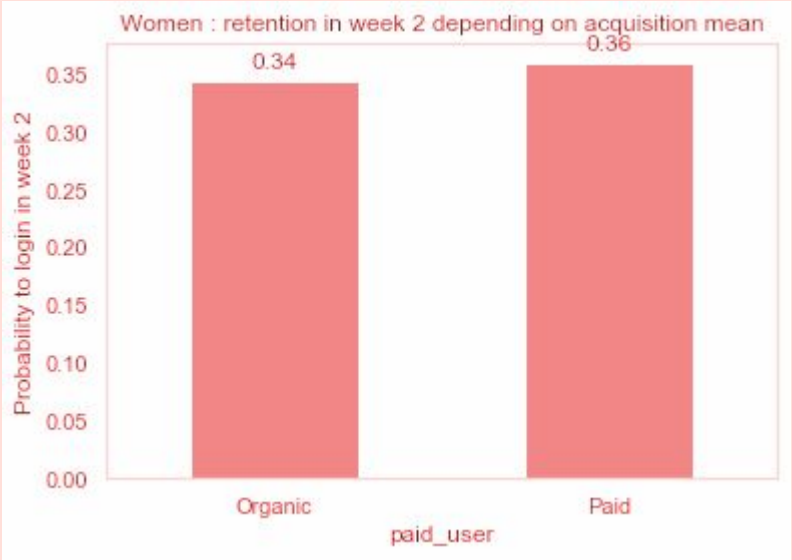



The more frequently user are connecting during the first 5 days, the more likely they are to stay in the app in week 2.

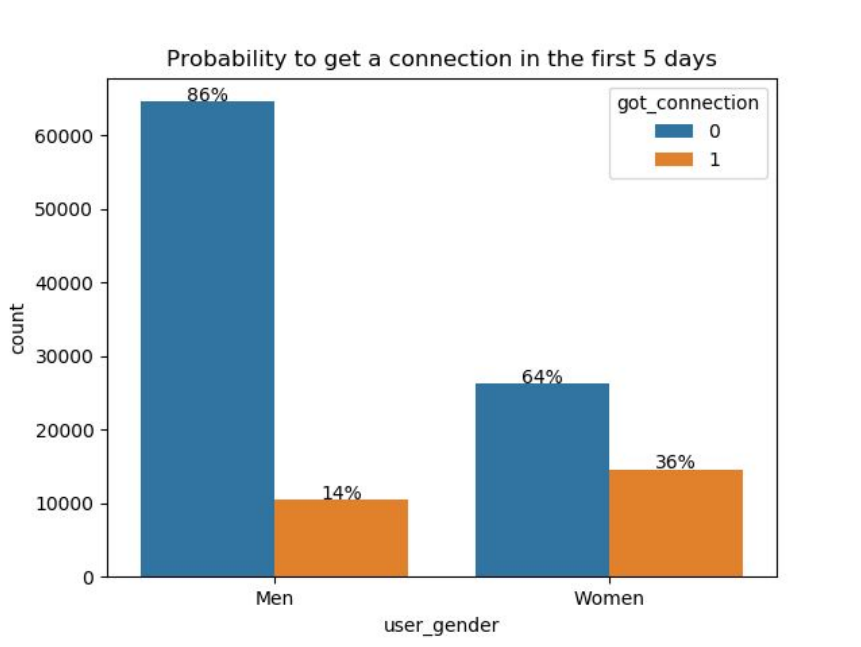
# IMPACT OF ACQUISITION MEAN ON CHURN



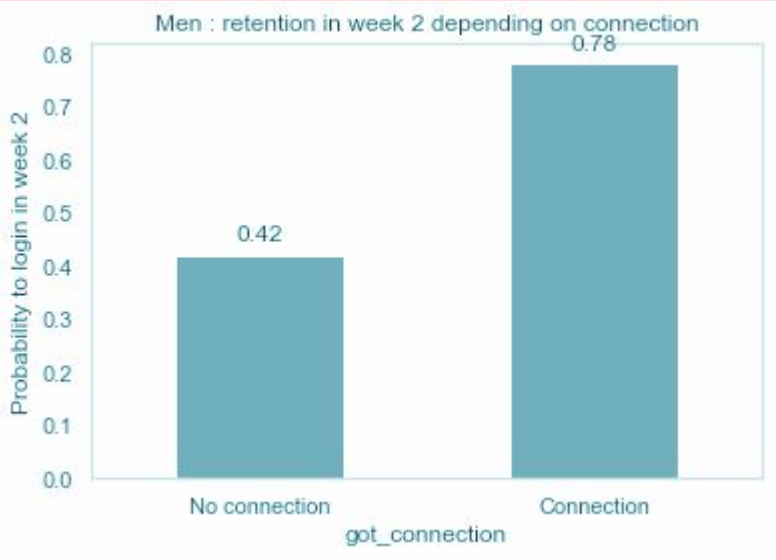

There does not seem to be a correlation between churn and acquisition mean

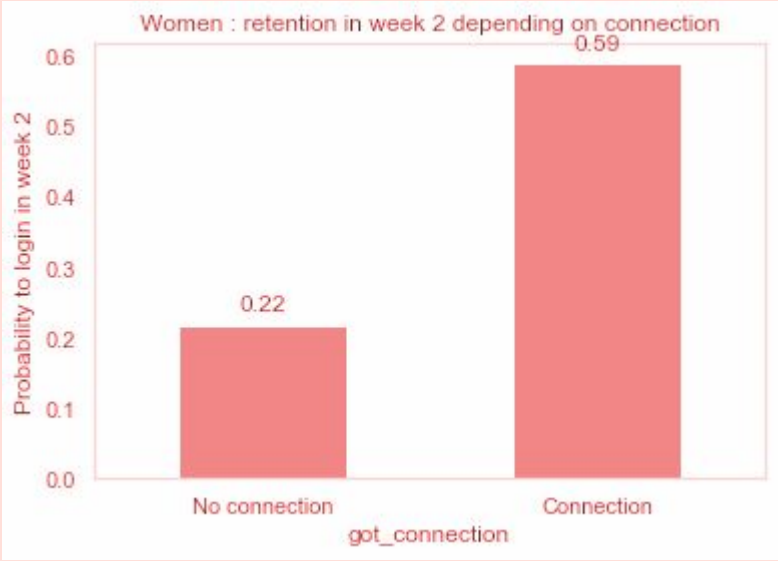
# CONNECTIONS DURING THE FIRST 5 DAYS




Women are more likely to get a connection during the first 5 days.

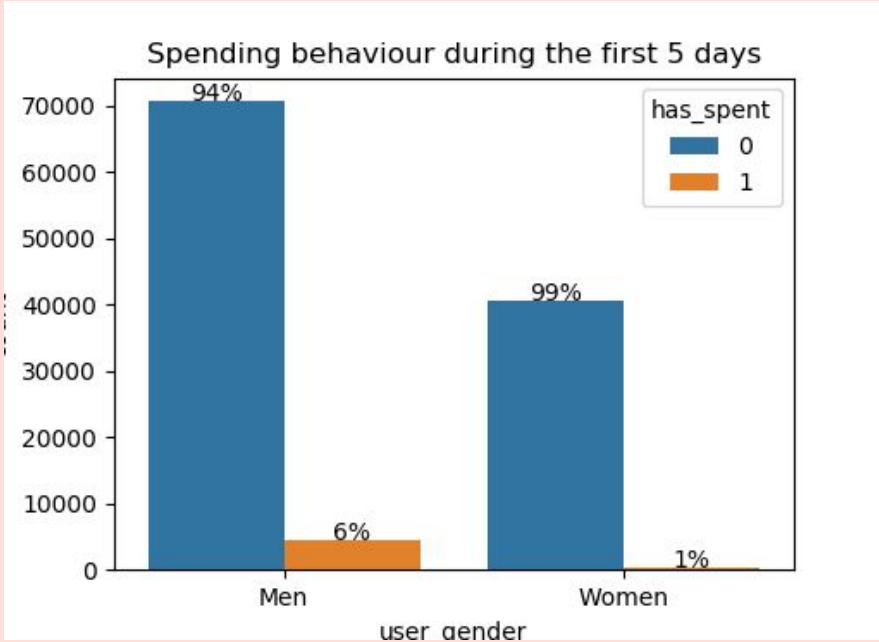
# IMPACT OF GETTING A CONNECTION ON CHURN



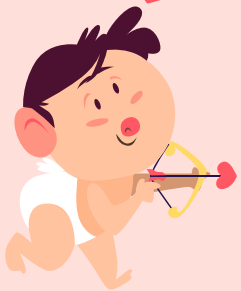




Getting a connection during the first 5 days has a significant impact on the probability of users to keep on using the app.

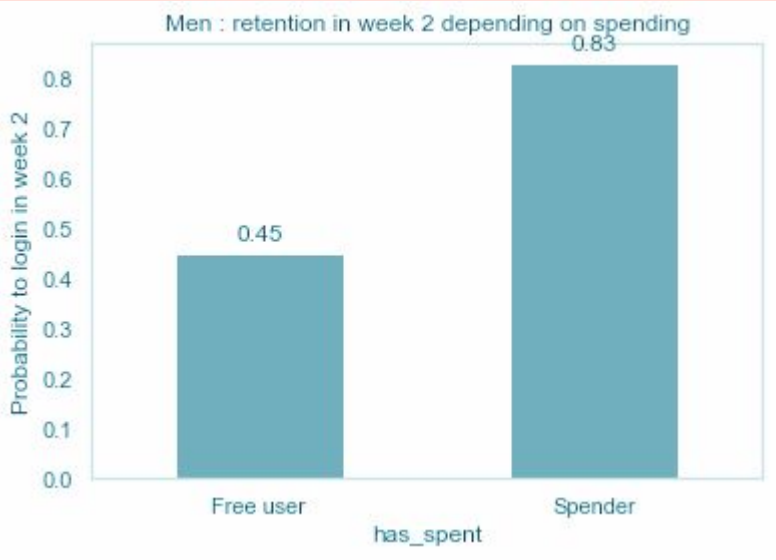
# PROPORTION OF SPENDERS DURING THE FIRST 5 DAYS

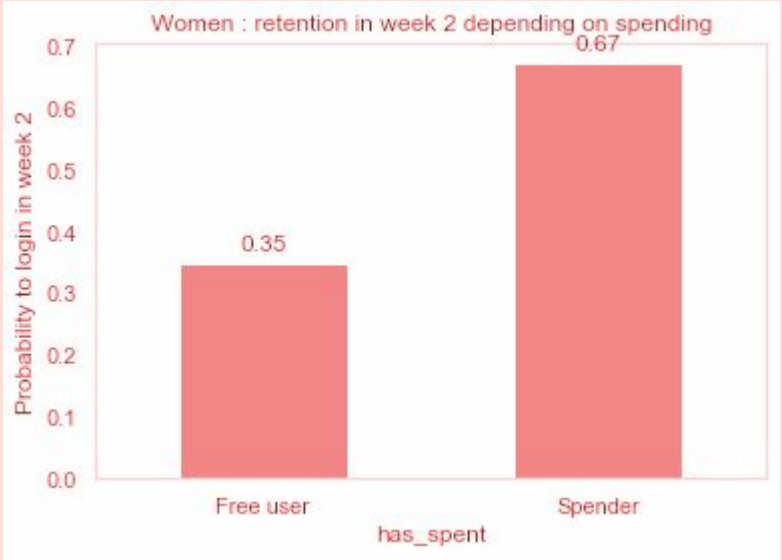



7% of the users are spending during the first 5 days, they are mainly men.



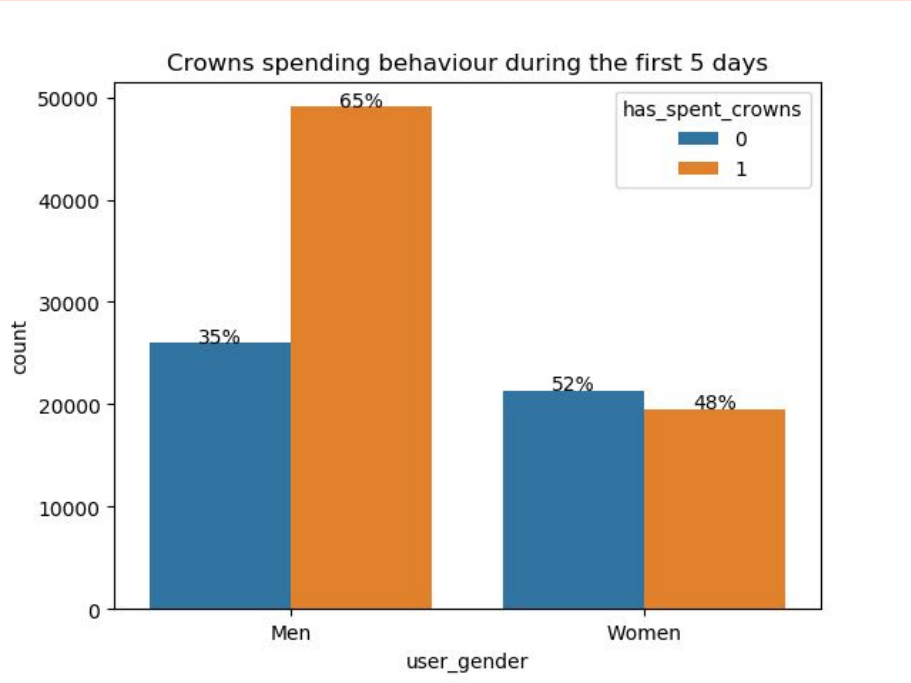
# IMPACT OF SPENDING ON CHURN



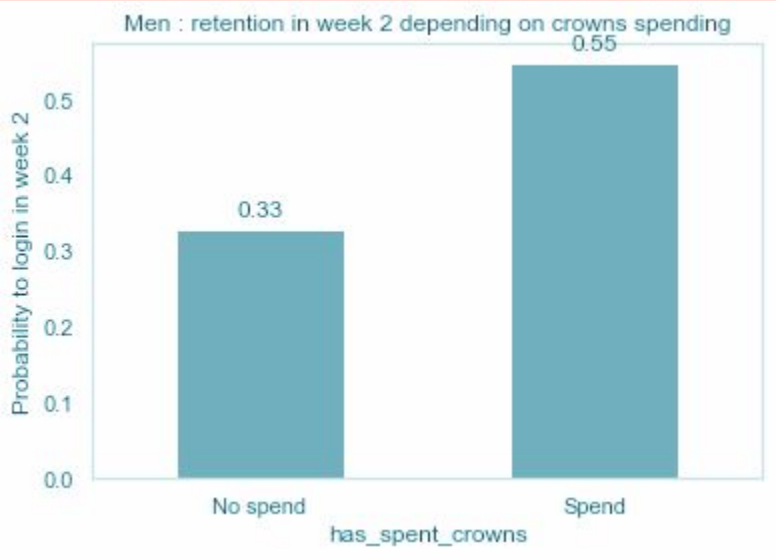

Spending is a retention driver, people who are investing money in the app are also more likely to invest time.

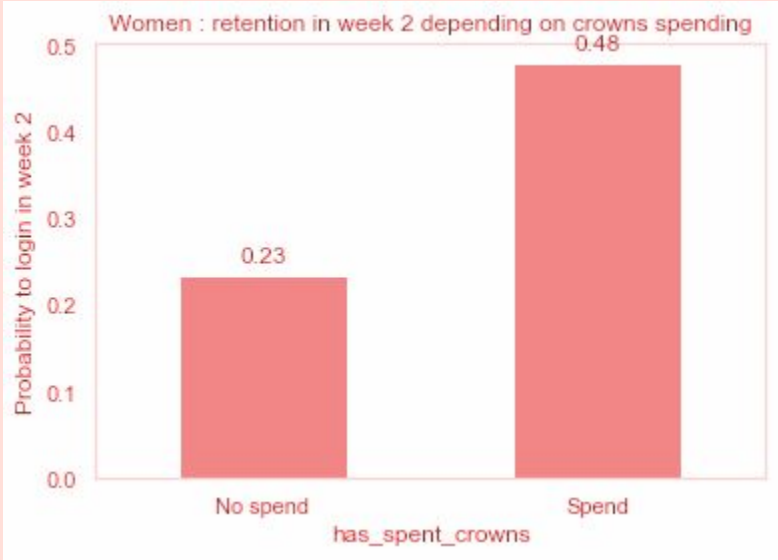
# PROPORTION OF VIRTUAL CURRENCY SPENDERS DURING THE FIRST 5 DAYS



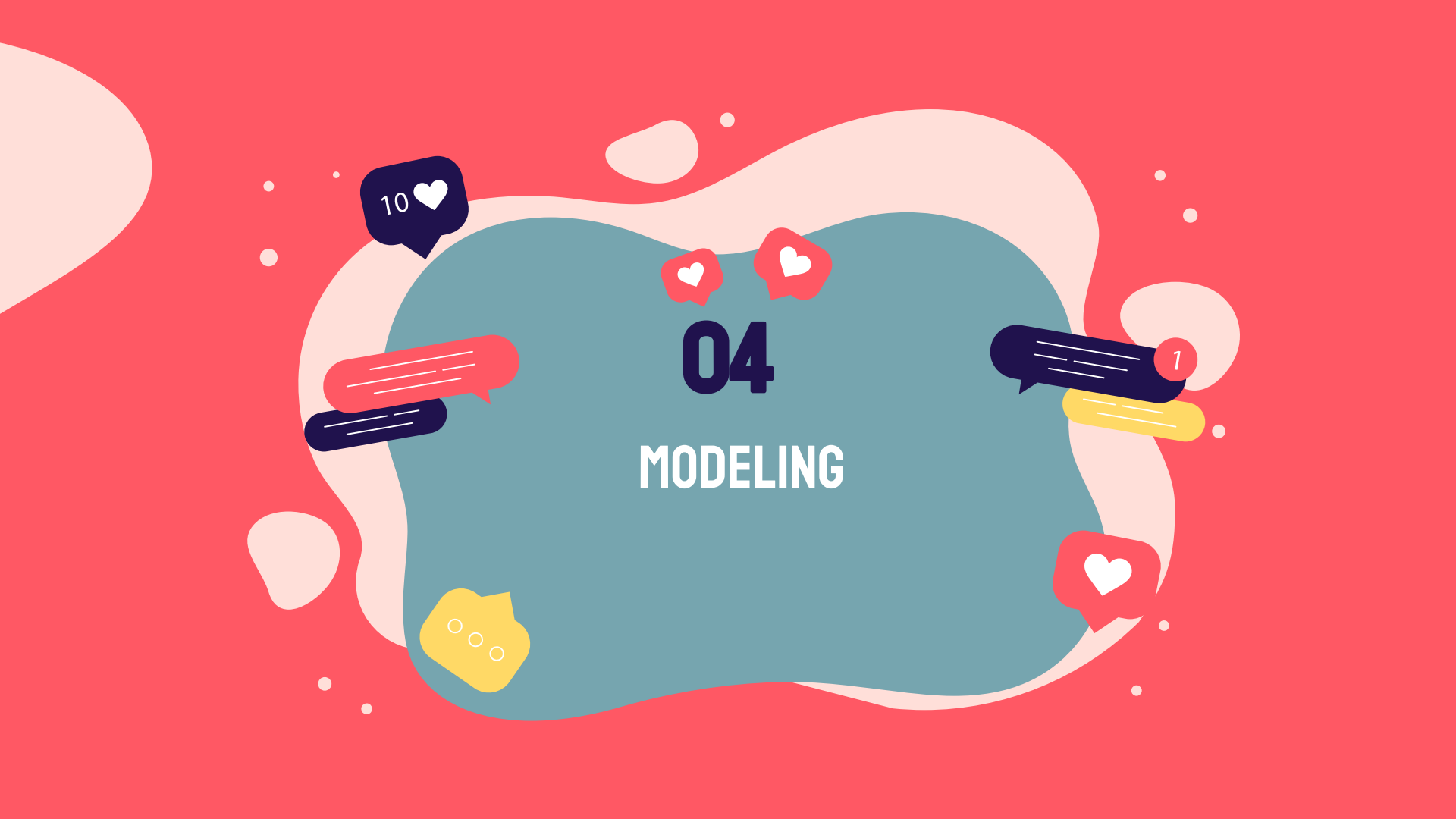

While half of the women are spending virtual currency during the first 5 days, the proportion rises to 65% for men, due to the economic model of the app

# IMPACT OF VIRTUAL CURRENCY SPENDING ON CHURN




Spending virtual currency generates better retention




# 04

# MODELING

# PROCESS STEPS

## Feature Engineering

### Logistic Regression with statsmodel

Remove variables with p-value > 5%

### RFE PCA

RFE : 15 features

PCA : 90% explained variance ratio
=> 11 features

## Model selection / training / testing

### Models

3 variations each :

- ★ Default
- ★ RFE
- ★ PCA

Logistic Regression
Decision Tree
Random Forest
KNeighbors
Naive Bayes
XGBoost
CatBoost

## Model validation

### Cross-validation & Hyper-parameter tuning

2 selected most performing models

Score *: precision*

Cross validation : *ReapetedStratifiedKFlod*

Hyper-parameter tuning : *GridSearchCV*

### Ensemble

2 most performing models

Process performed twice : for men and women

# MEN

# LOGISTIC REGRESSION WITH STATSMODEL

Results: Logit

| | | | |
|---|---|---|---|
| Model: | Logit | Pseudo R-squared: | 0.402 |
| Dependent Variable: | w2 | AIC: | 62183.6969 |
| Date: | 2020-03-05 15:35 | BIC: | 62534.3098 |
| No. Observations: | 75106 | Log-Likelihood: | -31054. |
| Df Model: | 37 | LL-Null: | -51936. |
| Df Residuals: | 75068 | LLR p-value: | 0.0000 |
| Converged: | 0.0000 | Scale: | 1.0000 |
| No. Iterations: | 35.0000 | | |

| | Coef. | Std.Err. | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| crowns_usd_5d | 0.3565 | 0.1160 | 3.0741 | 0.0021 | 0.1292 | 0.5838 |
| sub_usd_5d | 0.4104 | 0.0773 | 5.3091 | 0.0000 | 0.2589 | 0.5619 |
| discount_usd_5d | 17.6160 | 2502.7199 | 0.0070 | 0.9944 | -4887.6249 | 4922.8569 |
| chat_request_received_5d | 0.0257 | 0.0136 | 1.8870 | 0.0592 | -0.0010 | 0.0524 |
| chat_request_sent_5d | -0.0870 | 0.0264 | -3.2979 | 0.0010 | -0.1387 | -0.0353 |
| crowns_spent_chat_5d | 0.0068 | 0.0702 | 0.0965 | 0.9231 | -0.1308 | 0.1444 |
| crowns_spent_message_5d | -0.0662 | 0.0302 | -2.1908 | 0.0285 | -0.1254 | -0.0070 |
| crowns_spent_match_now_5d | -0.1794 | 0.0292 | -6.1520 | 0.0000 | -0.2366 | -0.1223 |
| crowns_spent_discover_5d | 0.0307 | 0.1122 | 0.2734 | 0.7845 | -0.1892 | 0.2505 |
| crowns_spent_pick_5d | -0.2096 | 0.0359 | -5.8314 | 0.0000 | -0.2800 | -0.1391 |
| crowns_spent_instant_match_now_5d | -0.0146 | 0.0375 | -0.3889 | 0.6973 | -0.0880 | 0.0589 |
| crowns_spent_more_pick_5d | -0.1387 | 0.0642 | -2.1603 | 0.0308 | -0.2646 | -0.0129 |
| rating_given | -0.1205 | 0.0108 | -11.1179 | 0.0000 | -0.1417 | -0.0992 |
| avg_rating_given | -0.0275 | 0.0113 | -2.4308 | 0.0151 | -0.0496 | -0.0053 |
| rating_received | 0.0428 | 0.0151 | 2.8314 | 0.0046 | 0.0132 | 0.0724 |
| avg_rating_received | -0.0776 | 0.0132 | -5.8737 | 0.0000 | -0.1035 | -0.0517 |
| login_5d | 1.9237 | 0.0140 | 137.0302 | 0.0000 | 1.8962 | 1.9512 |
| user_platform_ios | 0.0663 | 0.0230 | 2.8781 | 0.0040 | 0.0212 | 0.1115 |
| user_platform_web | 0.2006 | 0.0478 | 4.1970 | 0.0000 | 0.1069 | 0.2943 |
| country_BE | -0.3865 | 0.0732 | -5.2832 | 0.0000 | -0.5299 | -0.2431 |
| country_BR | -0.5796 | 0.0654 | -8.8606 | 0.0000 | -0.7078 | -0.4514 |
| country_CA | -0.3534 | 0.1276 | -2.7681 | 0.0056 | -0.6035 | -0.1032 |
| country_CH | -0.3841 | 0.0771 | -4.9818 | 0.0000 | -0.5353 | -0.2330 |
| country_DE | -0.5500 | 0.0427 | -12.8662 | 0.0000 | -0.6337 | -0.4662 |
| country_FR | -0.3248 | 0.0249 | -13.0297 | 0.0000 | -0.3736 | -0.2759 |
| country_GB | -0.5537 | 0.0883 | -6.2709 | 0.0000 | -0.7267 | -0.3806 |
| country_IT | -0.4037 | 0.0348 | -11.5991 | 0.0000 | -0.4719 | -0.3355 |
| country_NL | -0.4938 | 0.1278 | -3.8641 | 0.0001 | -0.7443 | -0.2433 |
| country_US | -0.6101 | 0.0770 | -7.9224 | 0.0000 | -0.7611 | -0.4592 |
| country_other | -0.1115 | 0.2074 | -0.5373 | 0.5910 | -0.5180 | 0.2951 |
| country_other Africa | 0.0591 | 0.1833 | 0.3221 | 0.7474 | -0.3003 | 0.4184 |
| country_other Asia | -0.6247 | 0.0893 | -6.9981 | 0.0000 | -0.7997 | -0.4497 |
| country_other EU | -0.4003 | 0.0629 | -6.3630 | 0.0000 | -0.5237 | -0.2770 |
| country_other South America | -0.4847 | 0.1350 | -3.5910 | 0.0003 | -0.7493 | -0.2202 |
| age_29-38 | 0.0715 | 0.0245 | 2.9130 | 0.0036 | 0.0234 | 0.1196 |
| age_38-48 | 0.0509 | 0.0350 | 1.4529 | 0.1463 | -0.0178 | 0.1196 |
| age_more than 48 | 0.1895 | 0.0477 | 3.9708 | 0.0001 | 0.0960 | 0.2830 |
| connection | 0.3722 | 0.0368 | 10.1159 | 0.0000 | 0.3001 | 0.4444 |

**8 features removed**

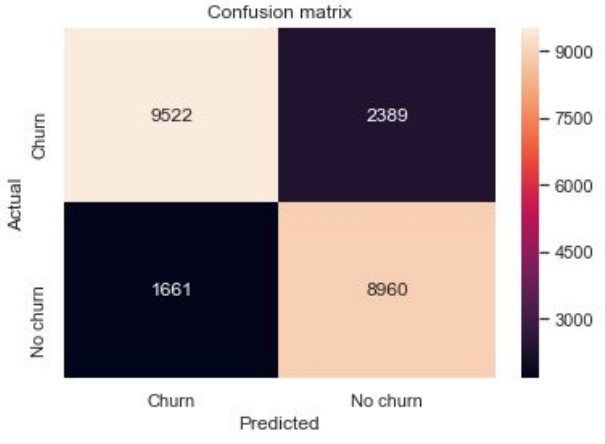
**Number of login days in the first 5 days has the highest coef**

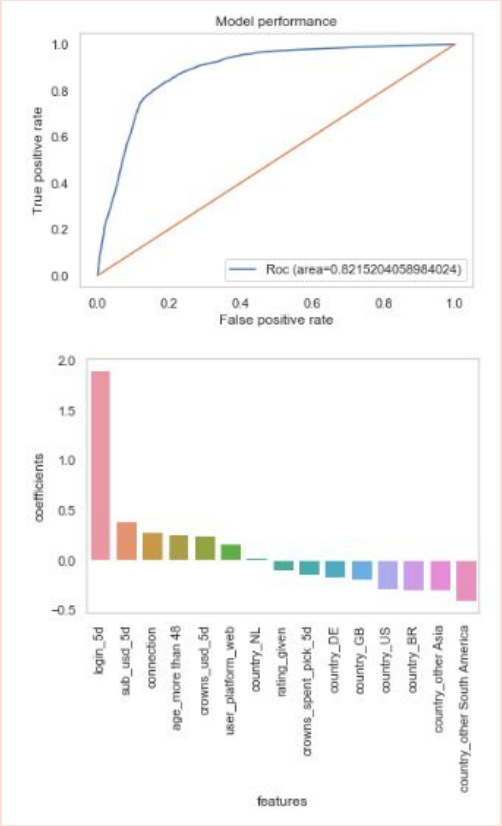
# LOGISTIC REGRESSION WITH RFE

Classification report :

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.85 | 0.80 | 0.82 | 11911 |
| 1 | 0.79 | 0.84 | 0.82 | 10621 |
| accuracy | | | 0.82 | 22532 |
| macro avg | 0.82 | 0.82 | 0.82 | 22532 |
| weighted avg | 0.82 | 0.82 | 0.82 | 22532 |

Accuracy Score : 0.820255636428191
Precision Score : 0.7894968719710987
Area under curve : 0.8215204058984024



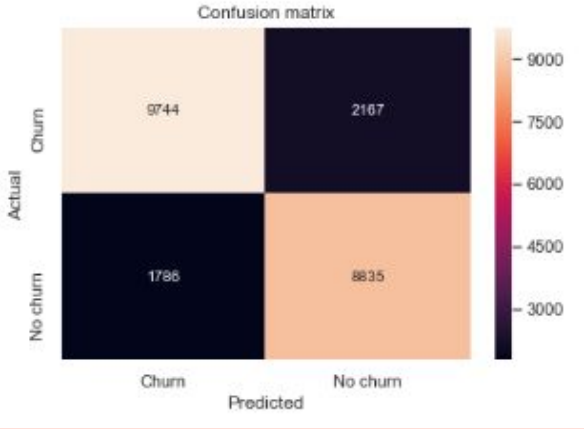

**Precision score : 79%**

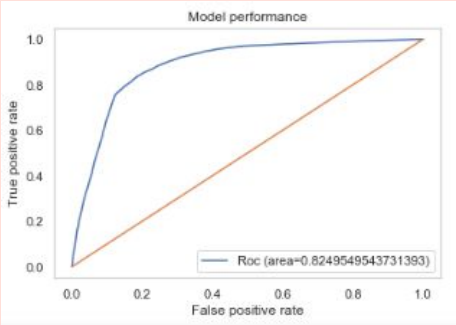
**Most important coefficients :**

- ★ **Number of login days**
- ★ **Purchase of a subscription**
- ★ **Got a connection**

# CATBOOST WITH RFE

```
Classification report :
              precision    recall  f1-score   support

           0       0.85      0.82      0.83     11911
           1       0.80      0.83      0.82     10621

    accuracy                           0.82     22532
   macro avg       0.82      0.82      0.82     22532
weighted avg       0.83      0.82      0.82     22532

Accuracy   Score :  0.8245606248890467
Precision Score :  0.8030358116706053
Area under curve :  0.8249549543731393
```




**Precision score : 80.3%**

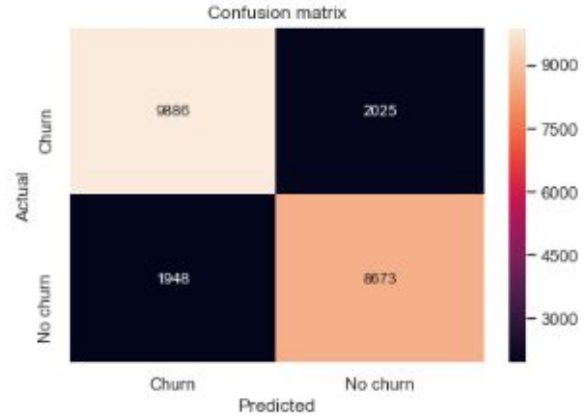
# CROSS VALIDATION & HYPER PARAMETER TUNING

**Impact of cross validation and hyper-parameter tuning on precision score**

| | Before | After |
|---|---|---|
| Logistic regression | 79% | 80% |
| Catboost | 80% | 80% |

# ENSEMBLE

```
Classification report :
              precision    recall  f1-score   support

           0       0.84      0.83      0.83     11911
           1       0.81      0.82      0.81     10621

    accuracy                           0.82     22532
   macro avg       0.82      0.82      0.82     22532
weighted avg       0.82      0.82      0.82     22532

Accuracy   Score :  0.8236729984022724
Precision Score :  0.8107122826696579

Text(0.5, 1, 'Confusion matrix')
```




**Precision score : 81%**

**=> +0.7% compared to best model (Catboost)**

# WOMEN

# LOGISTIC REGRESSION WITH STATSMODEL

Results: Logit

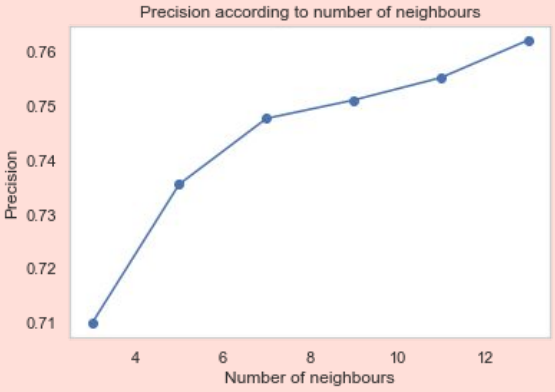
| | | | |
|---|---|---|---|
| Model: | Logit | Pseudo R-squared: | 0.391 |
| Dependent Variable: | w2 | AIC: | 32332.9777 |
| Date: | 2020-03-05 22:52 | BIC: | 32651.8593 |
| No. Observations: | 40881 | Log-Likelihood: | -16129. |
| Df Model: | 36 | LL-Null: | -26497. |
| Df Residuals: | 40844 | LLR p-value: | 0.0000 |
| Converged: | 1.0000 | Scale: | 1.0000 |
| No. Iterations: | 7.0000 | | |

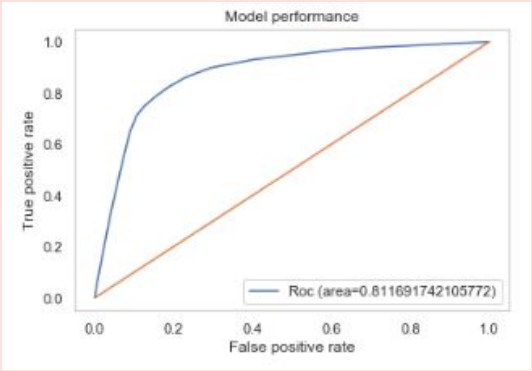
| | Coef. | Std.Err. | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| crowns_usd_5d | 0.2540 | 0.2516 | 1.0093 | 0.3128 | -0.2392 | 0.7472 |
| sub_usd_5d | 0.5096 | 0.1780 | 2.8630 | 0.0042 | 0.1607 | 0.8585 |
| chat_request_received_5d | 0.0892 | 0.0191 | 4.6694 | 0.0000 | 0.0517 | 0.1266 |
| chat_request_sent_5d | 0.1164 | 0.0170 | 6.8556 | 0.0000 | 0.0831 | 0.1497 |
| crowns_spent_chat_5d | 0.2186 | 0.3206 | 0.6816 | 0.4955 | -0.4099 | 0.8470 |
| crowns_spent_message_5d | 0.0992 | 0.0421 | 2.3574 | 0.0184 | 0.0167 | 0.1816 |
| crowns_spent_match_now_5d | -0.1558 | 0.0353 | -4.4086 | 0.0000 | -0.2250 | -0.0865 |
| crowns_spent_discover_5d | 0.5963 | 1.1144 | 0.5350 | 0.5926 | -1.5879 | 2.7804 |
| crowns_spent_pick_5d | -0.0148 | 0.0453 | -0.3272 | 0.7435 | -0.1037 | 0.0740 |
| crowns_spent_instant_match_now_5d | 0.0193 | 0.0357 | 0.5421 | 0.5877 | -0.0505 | 0.0892 |
| crowns_spent_more_pick_5d | -0.1589 | 0.0443 | -3.5847 | 0.0003 | -0.2458 | -0.0720 |
| rating_given | -0.1123 | 0.0172 | -6.5308 | 0.0000 | -0.1461 | -0.0786 |
| avg_rating_given | -0.0552 | 0.0160 | -3.4585 | 0.0005 | -0.0864 | -0.0239 |
| rating_received | 0.1232 | 0.0219 | 5.6292 | 0.0000 | 0.0803 | 0.1660 |
| avg_rating_received | -0.1792 | 0.0205 | -8.7299 | 0.0000 | -0.2195 | -0.1390 |
| login_5d | 1.8103 | 0.0196 | 92.1469 | 0.0000 | 1.7718 | 1.8488 |
| user_platform_ios | 0.1198 | 0.0314 | 3.8160 | 0.0001 | 0.0582 | 0.1813 |
| user_platform_web | -0.0637 | 0.0630 | -1.0110 | 0.3120 | -0.1871 | 0.0598 |
| country_BE | -1.2530 | 0.1006 | -12.4557 | 0.0000 | -1.4502 | -1.0558 |
| country_BR | -1.4126 | 0.0834 | -16.9415 | 0.0000 | -1.5760 | -1.2492 |
| country_CA | -1.2905 | 0.2310 | -5.5873 | 0.0000 | -1.7432 | -0.8378 |
| country_CH | -1.1214 | 0.0984 | -11.3996 | 0.0000 | -1.3142 | -0.9286 |
| country_DE | -1.2337 | 0.0598 | -20.6277 | 0.0000 | -1.3509 | -1.1165 |
| country_FR | -1.1814 | 0.0334 | -35.3409 | 0.0000 | -1.2469 | -1.1159 |
| country_GB | -1.1809 | 0.1418 | -8.3276 | 0.0000 | -1.4588 | -0.9030 |
| country_IT | -1.0446 | 0.0506 | -20.6307 | 0.0000 | -1.1438 | -0.9453 |
| country_NL | -1.4666 | 0.1956 | -7.4981 | 0.0000 | -1.8500 | -1.0832 |
| country_US | -1.0914 | 0.1310 | -8.3304 | 0.0000 | -1.3482 | -0.8346 |
| country_other | -0.4836 | 0.2997 | -1.6140 | 0.1065 | -1.0710 | 0.1037 |
| country_other Africa | -1.3847 | 0.2983 | -4.6416 | 0.0000 | -1.9695 | -0.8000 |
| country_other Asia | -0.9005 | 0.1515 | -5.9444 | 0.0000 | -1.1974 | -0.6036 |
| country_other EU | -1.0980 | 0.1124 | -9.7665 | 0.0000 | -1.3184 | -0.8777 |
| country_other South America | -1.2904 | 0.2674 | -4.8264 | 0.0000 | -1.8144 | -0.7664 |
| age_29-38 | 0.1297 | 0.0346 | 3.7502 | 0.0002 | 0.0619 | 0.1976 |
| age_38-48 | 0.1541 | 0.0413 | 3.7331 | 0.0002 | 0.0732 | 0.2350 |
| age_more_than_48 | 0.4647 | 0.0481 | 9.6580 | 0.0000 | 0.3704 | 0.5590 |
| connection | -0.0008 | 0.0395 | -0.0193 | 0.9846 | -0.0781 | 0.0766 |

**7 features removed**

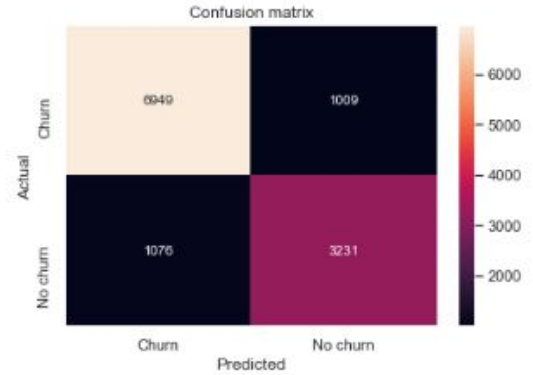
**Number of login days in the first 5 days has the highest coef**

# KNN WITH RFE




```
Classification report :
              precision    recall  f1-score   support

           0       0.87      0.87      0.87      7958
           1       0.76      0.75      0.76      4307

    accuracy                           0.83     12265
   macro avg       0.81      0.81      0.81     12265
weighted avg       0.83      0.83      0.83     12265

Accuracy   Score :  0.830004076640848
Precision Score :  0.7620283018867925
Area under curve :  0.811691742105772
```



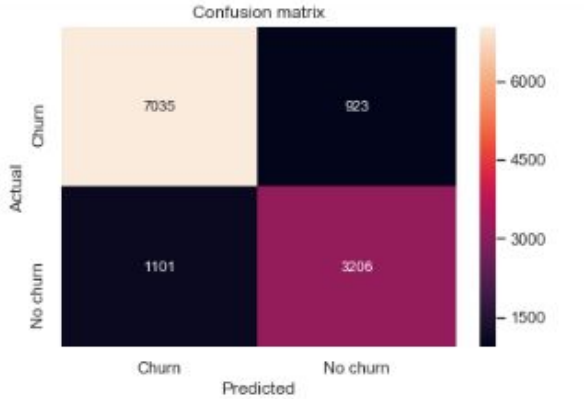

**Precision score : 76.2%**

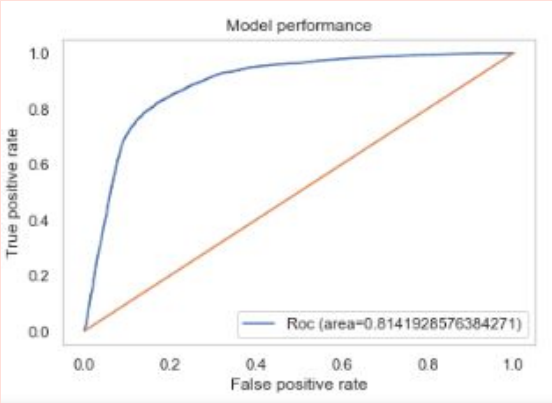
# CATBOOST WITH RFE

Classification report :

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.86 | 0.88 | 0.87 | 7958 |
| 1 | 0.78 | 0.74 | 0.76 | 4307 |
| accuracy | | | 0.83 | 12265 |
| macro avg | 0.82 | 0.81 | 0.82 | 12265 |
| weighted avg | 0.83 | 0.83 | 0.83 | 12265 |

Accuracy Score : 0.8349775784753363
Precision Score : 0.7764591910874303
Area under curve : 0.8141928576384271




**Precision score : 77.6%**

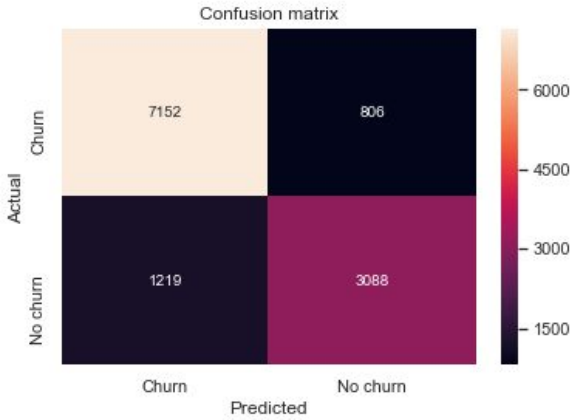
# CROSS VALIDATION & HYPER PARAMETER TUNING

**Impact of cross validation and hyper-parameter tuning on precision score**

| | Before | After |
|---|---|---|
| KNeighbors | 76.2% | 77.7% |
| Catboost | 77.6% | 78.4% |

# ENSEMBLE

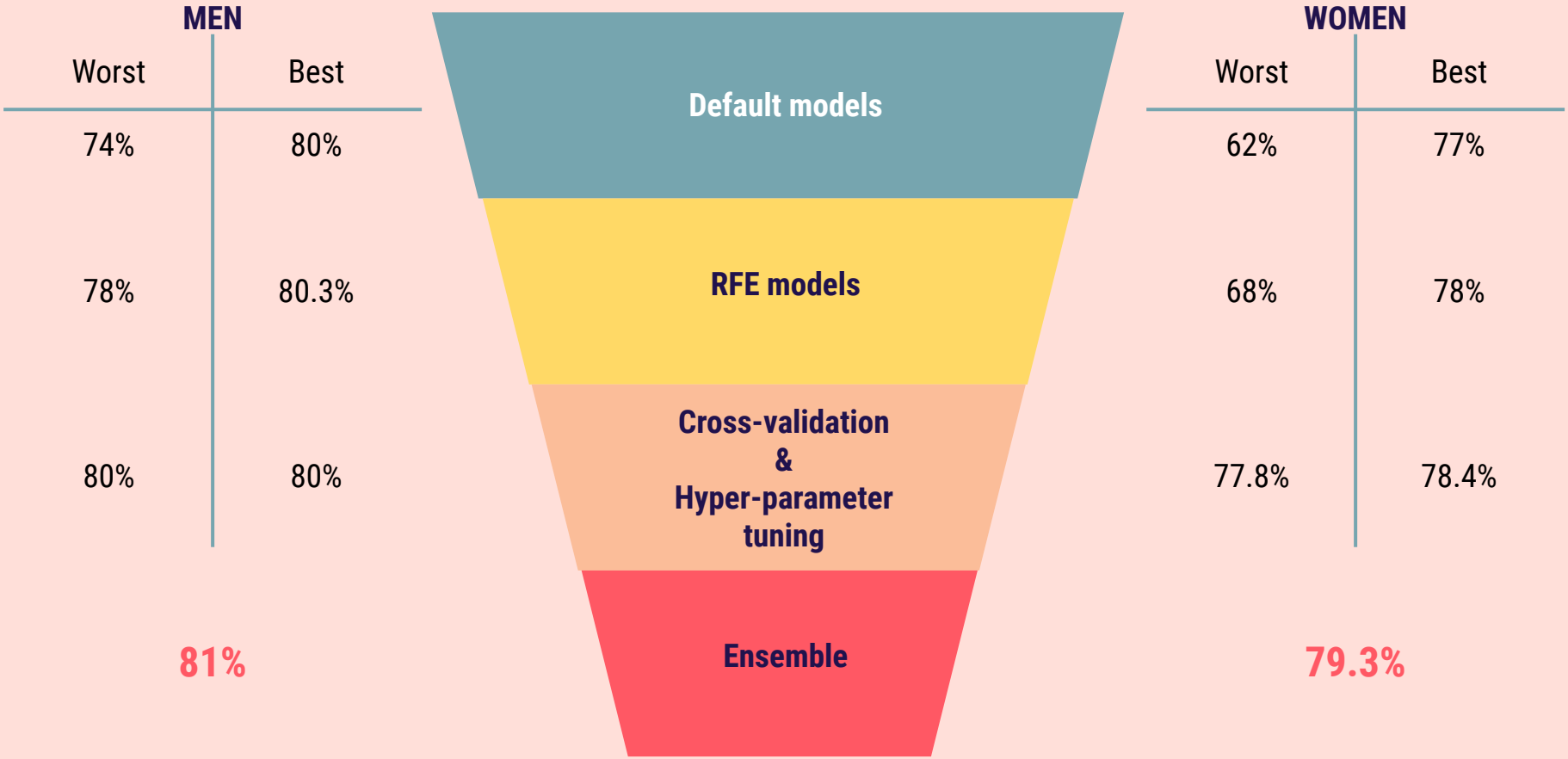
```
Classification report :
              precision    recall  f1-score   support

           0       0.85      0.90      0.88      7958
           1       0.79      0.72      0.75      4307

    accuracy                           0.83     12265
   macro avg       0.82      0.81      0.81     12265
weighted avg       0.83      0.83      0.83     12265

Accuracy   Score :  0.8348960456583775
Precision Score :  0.79301489470981

Text(0.5, 1, 'Confusion matrix')
```



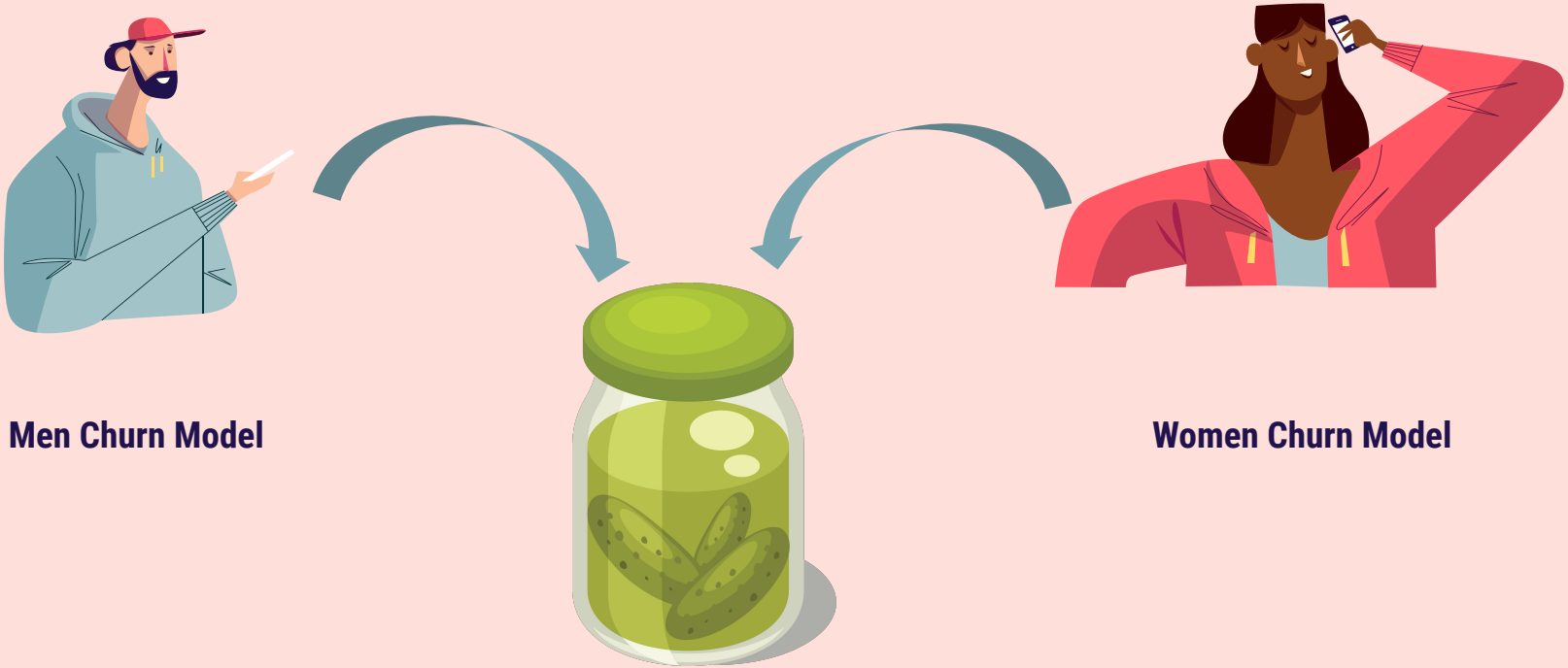

**Precision score : 79.3%**

**=> +0.9% compared to best model (Catboost)**

# SUMMARY


MEN
Worst
Best
74%
80%
78%
80.3%
80%
80%
81%
Default models
RFE models
Cross-validation
&
Hyper-parameter
tuning
Ensemble
WOMEN
Worst
Best
62%
77%
68%
78%
77.8%
78.4%
79.3%

# PICKLE



Men Churn Model

Women Churn Model

# CUSTOMER DELIVERABLE

## SQL QUERIES

Queries to run on a daily basis and to input to the python file that will perform the prediction

## MODEL

Best model for each gender

## PYTHON

Python file that will :

- ★ take the csv files as an input
- ★ clean and transform the data
- ★ apply the model
- ★ Generate a csv file with the list of predicted churners

# CONCLUSION

Churn can be predicted with a precision of :

- ★ **81%** for men
- ★ **79.3%** for women

Suggestion of action towards user who are identified as potential churners

- ★ Send targeted notifications
- ★ Offer virtual currency
- ★ Offer additional matchs

# THANKS!

Does anyone have any questions?