# Credit Sesame User Clustering and Product Offering Prediction

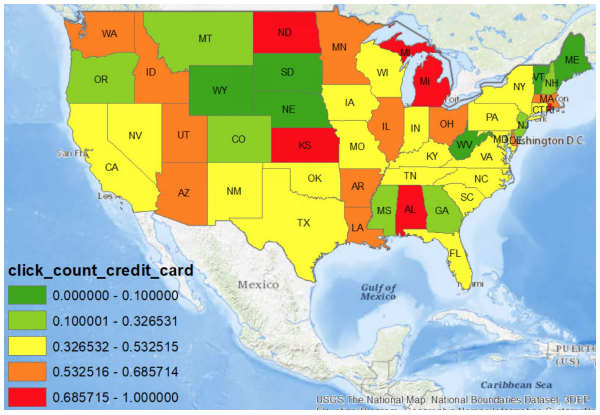
## Introduction

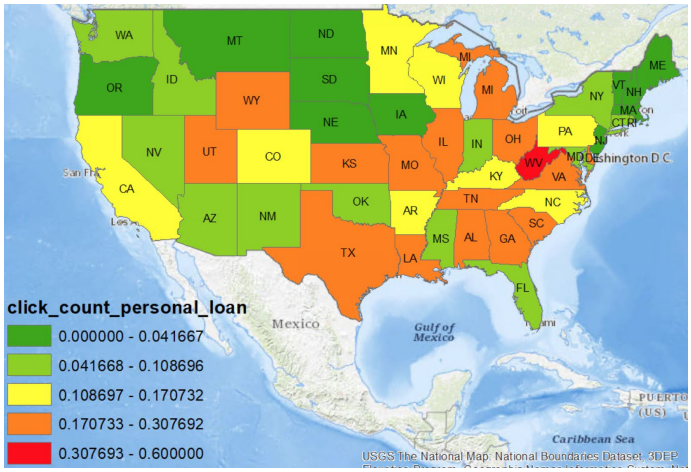
Clickstream data can reveal important information on user inclination for different products and provide valuable insights on orchestrating targeted offering and advertising. In this datathon project, we looked into how different demographic clusters respond to major product offerings such as credit cards & credit products and mortgages & loans.

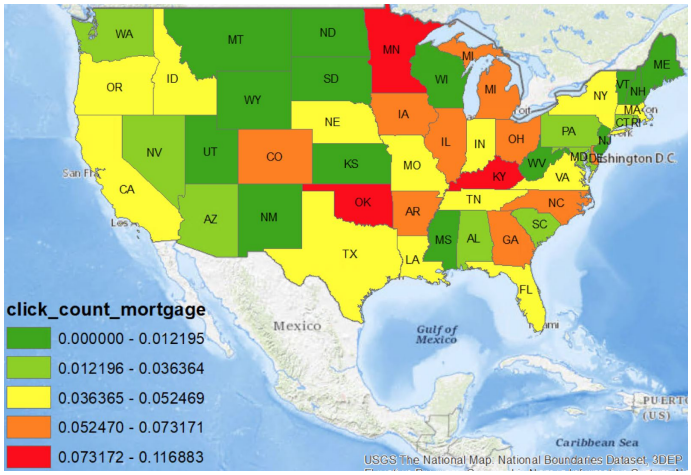
## Exploration

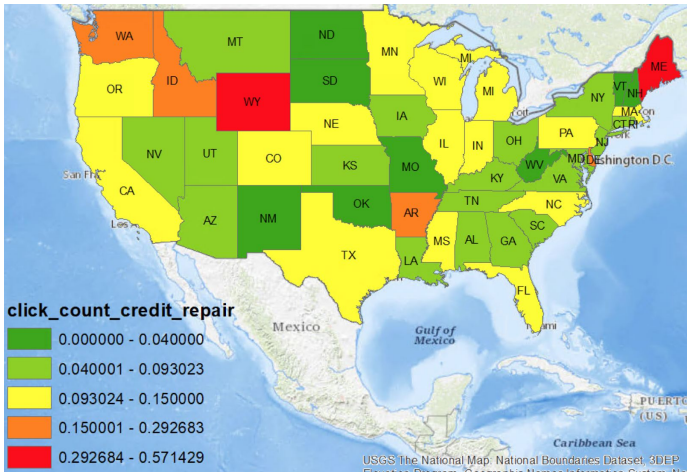
### 1. Geographical visualization of produce preferences

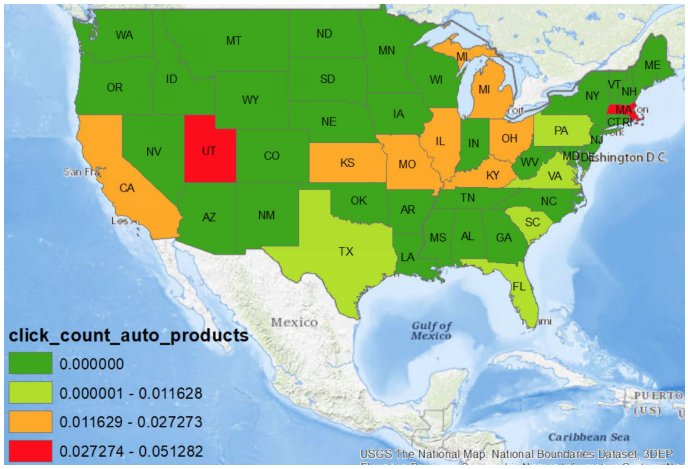
In this section, we draw the thematic map of user's click events in each state. These maps could show us which products people prefer in different states. Therefore, company could adjust their product recommendations for each state based on these results, such as notifying people who lived in ND, MI, KS and  AL which are most interested in credit card & products.

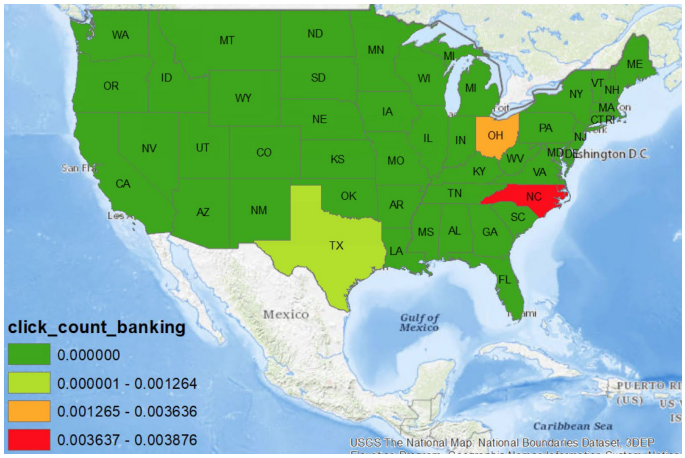



WA
MT
ND
MN
OR
ID
SD
WI
MI
WY
NE
IA
IL
IN
OH
PA
NY
VT
NH
ME
NV
UT
CO
KS
MO
WV
MD
VA
KY
CA
OK
AR
TN
NC
AZ
NM
SC
MS
AL
GA
TX
LA
FL
Mexico
Gulf of Mexico
Caribbean Sea
click_count_personal_loan
0.000000 - 0.041667
0.041668 - 0.108696
0.108697 - 0.170732
0.170733 - 0.307692
0.307693 - 0.600000
USGS The National Map: National Boundaries Dataset, 3DEP


WA
MT
ND
MN
OR
ID
SD
WI
MI
WY
NE
IA
IL
IN
OH
PA
NY
VT
NH
ME
NV
UT
CO
KS
MO
WV
MD
VA
KY
CA
OK
AR
TN
NC
AZ
NM
SC
MS
AL
GA
TX
LA
FL
Mexico
Gulf of Mexico
Caribbean Sea
click_count_mortgage
0.000000 - 0.012195
0.012196 - 0.036364
0.036365 - 0.052469
0.052470 - 0.073171
0.073172 - 0.116883
USGS The National Map: National Boundaries Dataset, 3DEP

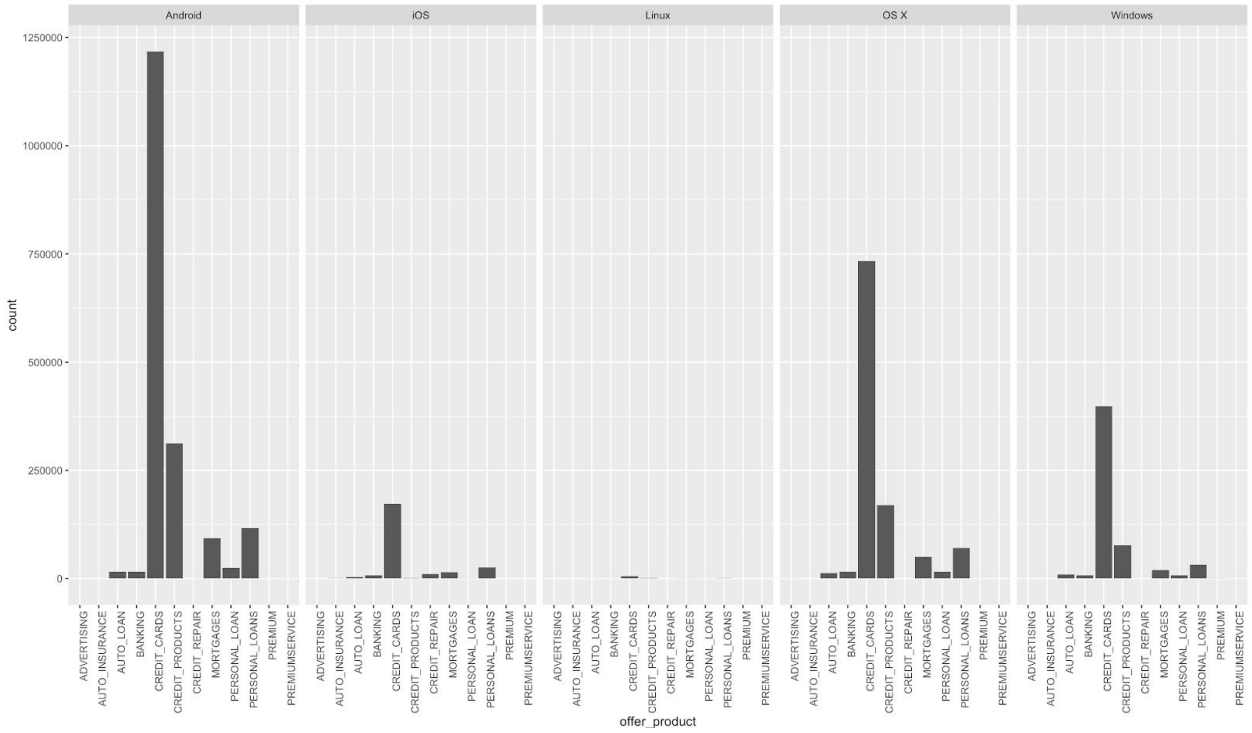


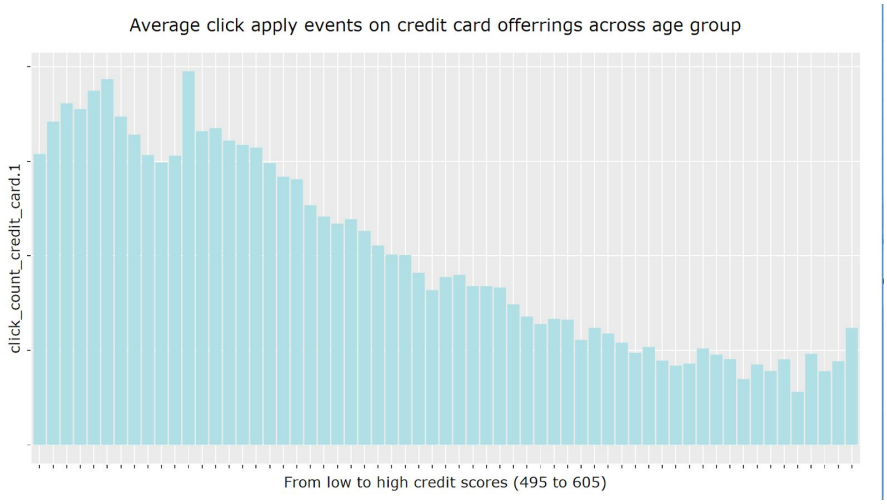

## 2. Product preferences for different devices

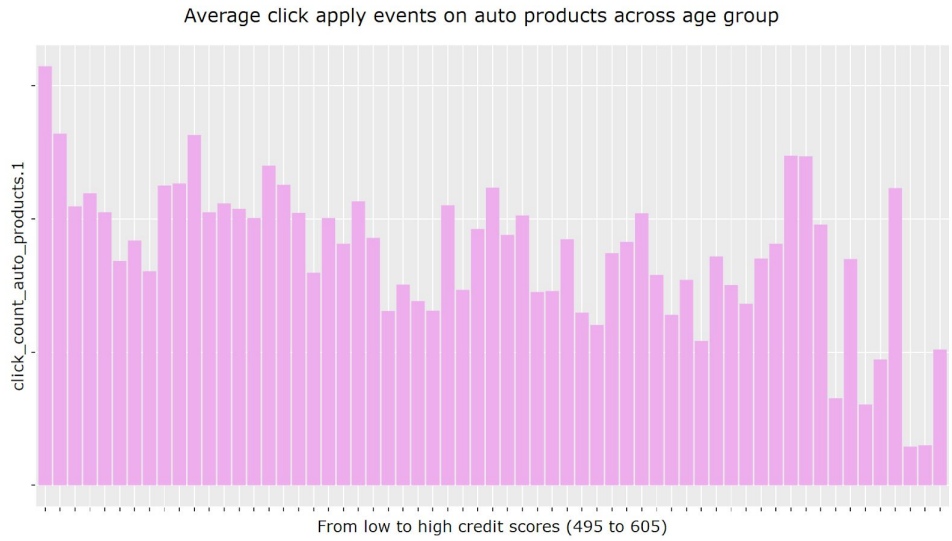
In this section, we draw the bar plots for different devices to show user's preference.



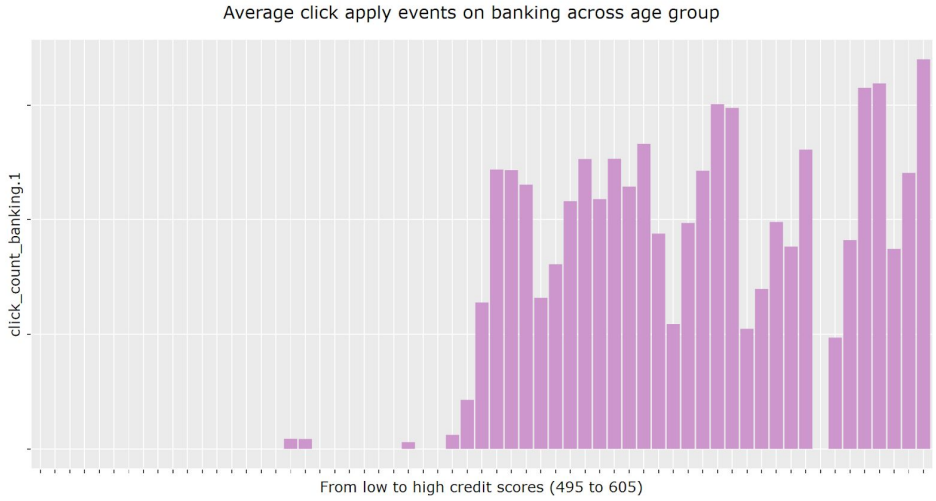

## 3. Product preferences on different credit score groups

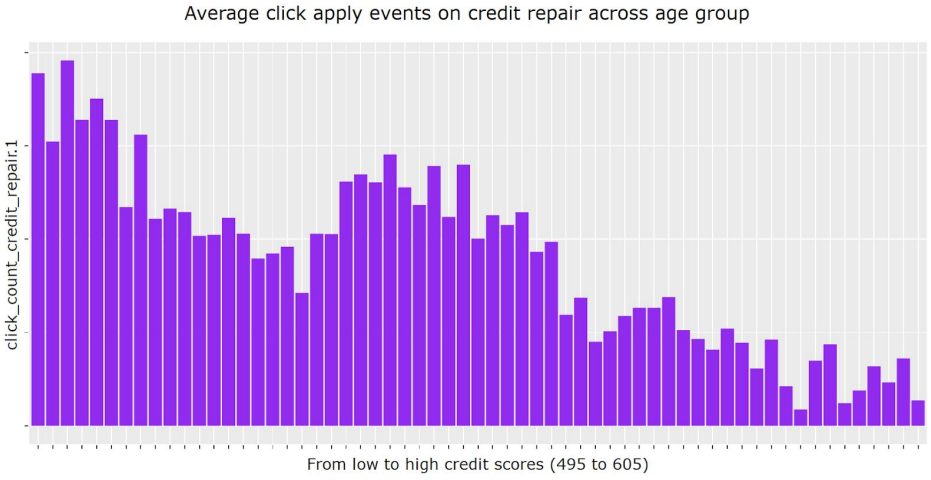
Lower customers' credit scores, they tend to click and apply for credit cards more frequently.

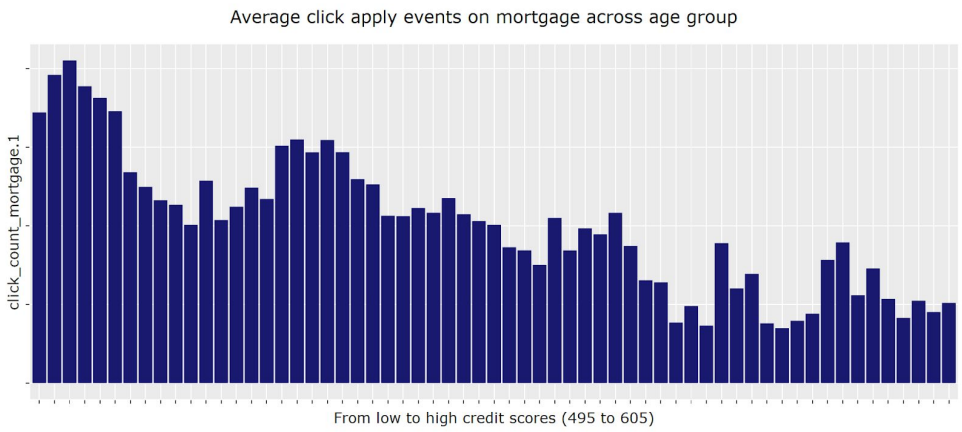


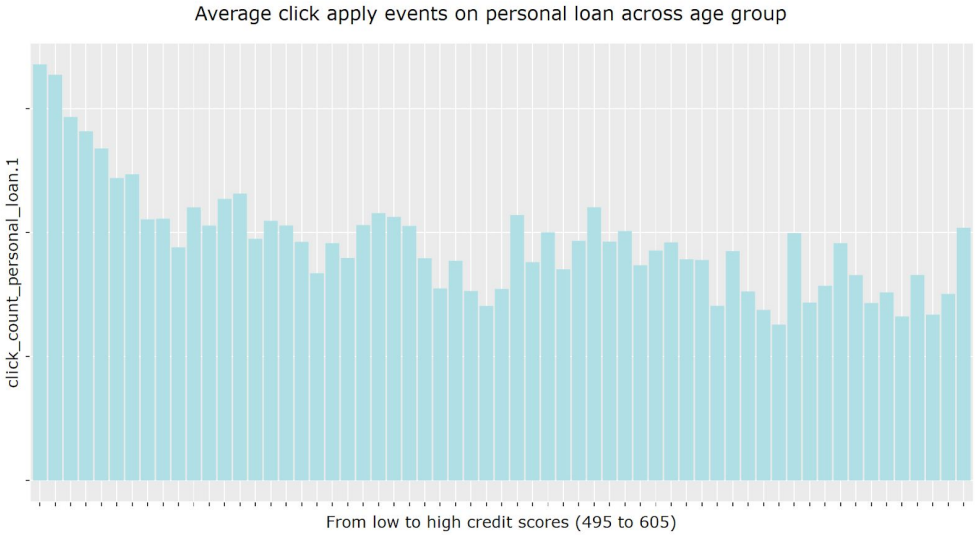



This an obviou

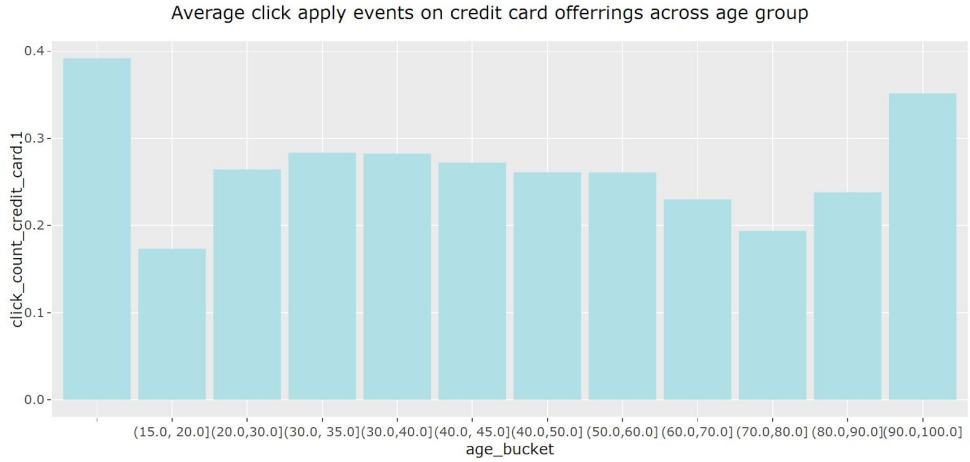


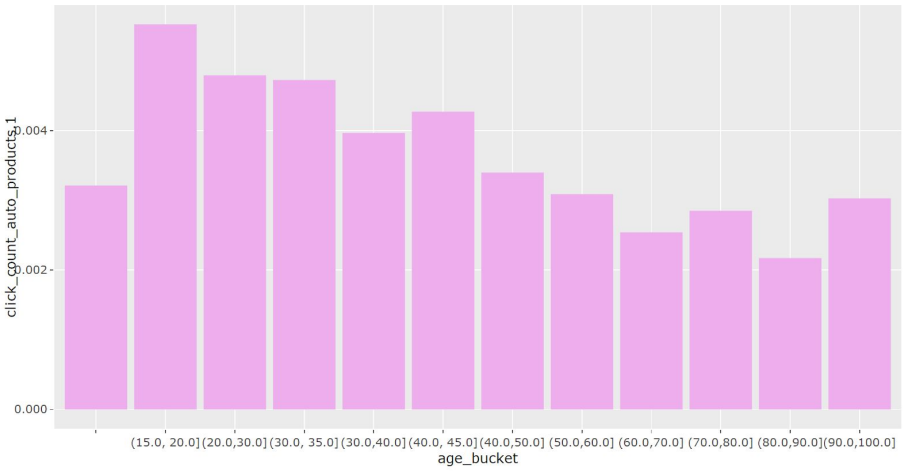


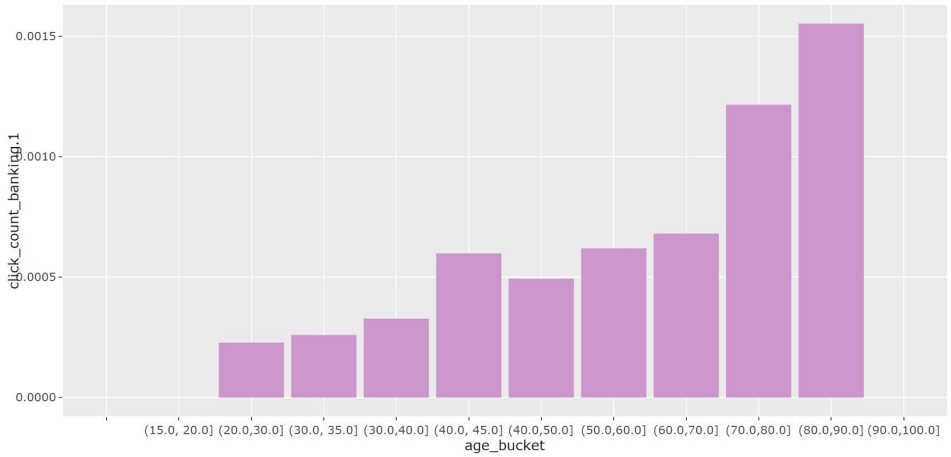

## 4. Product preferences on different age groups



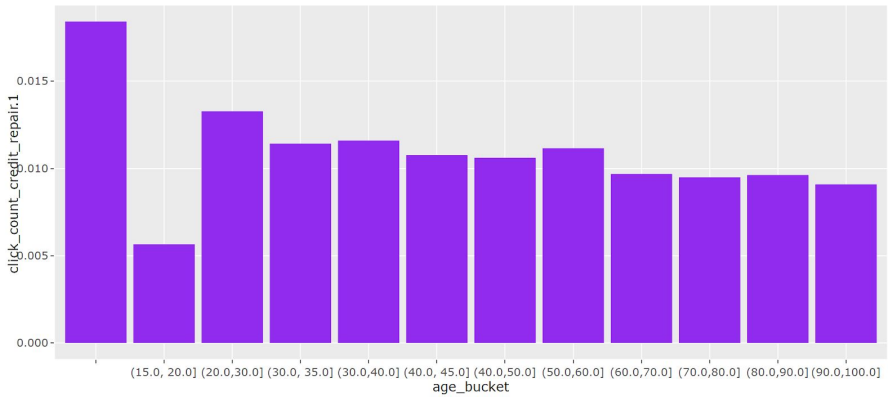
Average click apply events on auto products across age group

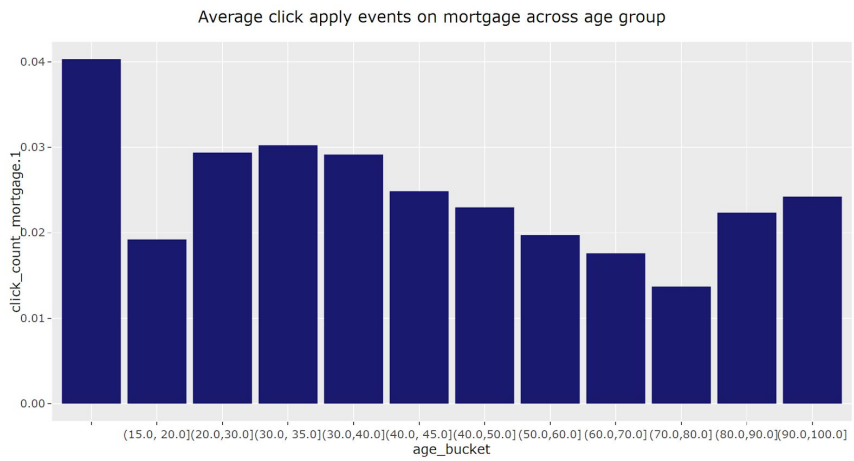


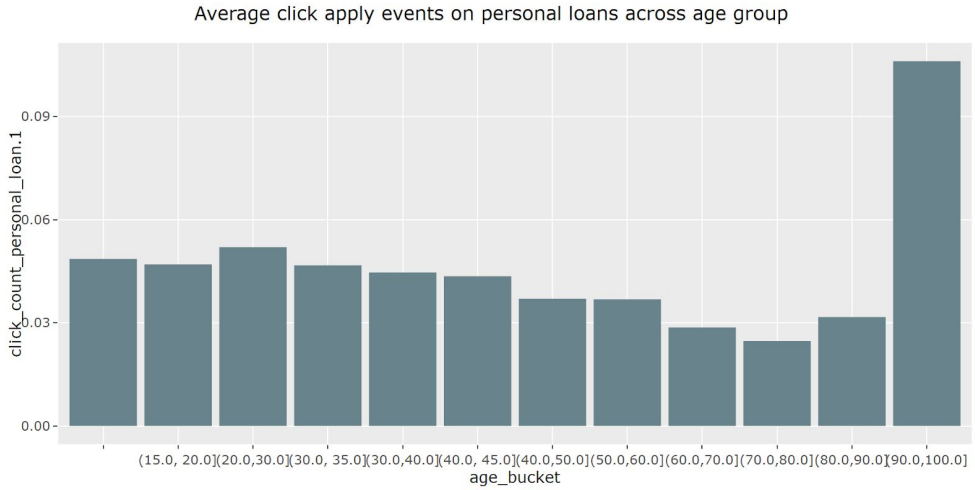

Average click apply events on banking across age group




Average click apply events on credit repair across age group

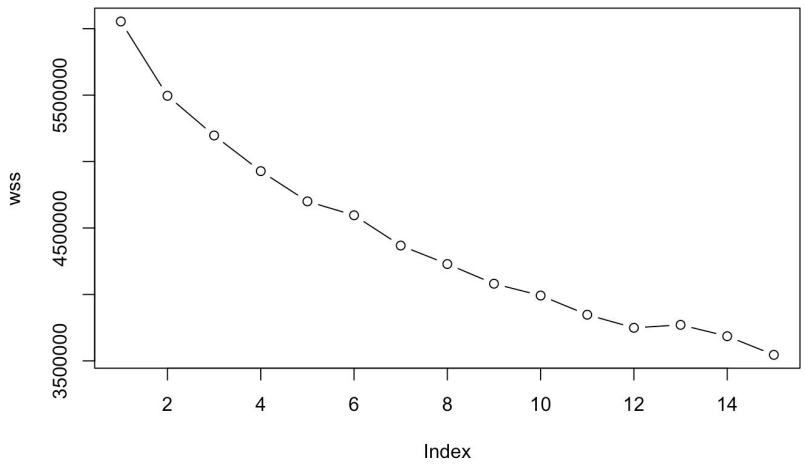




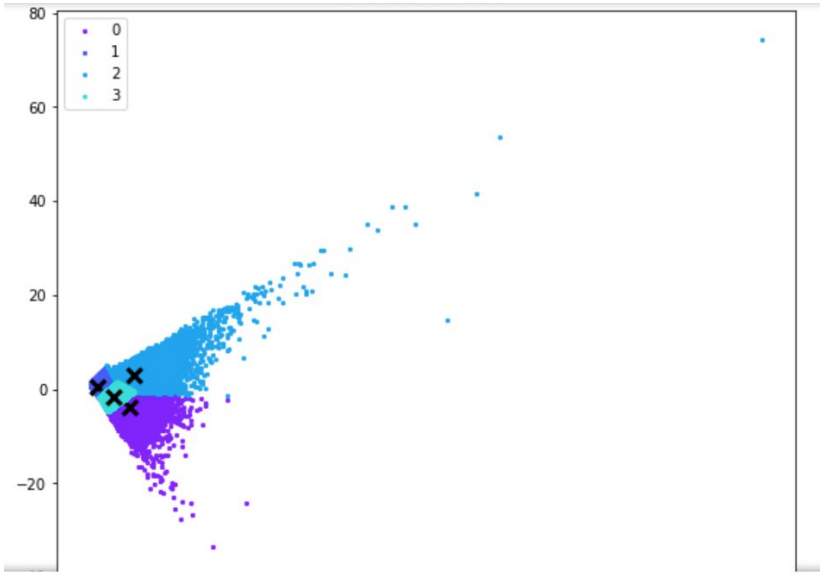




# Clustering

Using the numerical variables which are the majority of user profile records, we iterated on different K means clusters and decided to use 4 clusters for the best interpretability of the model output.



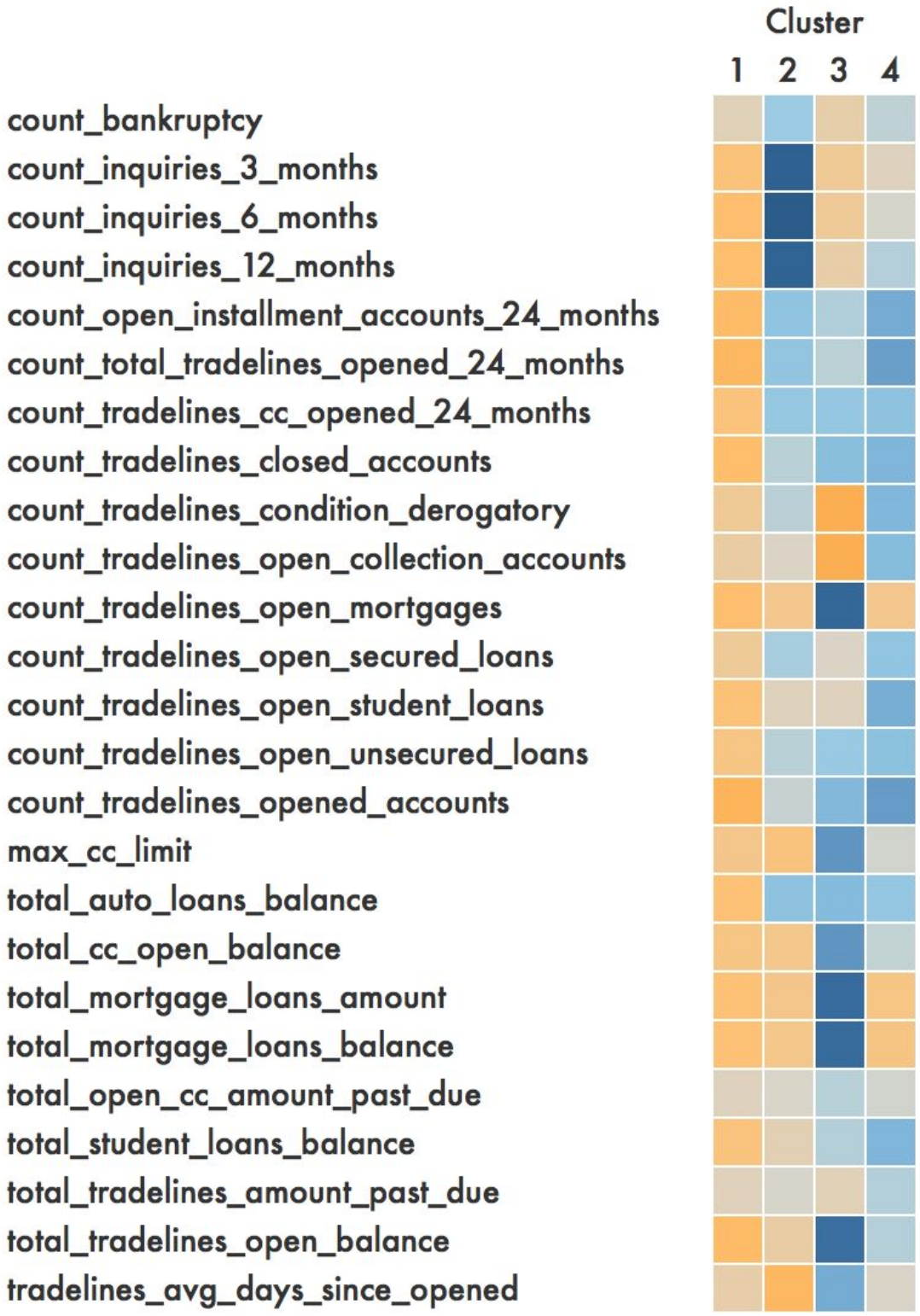

K-means result




**User profiling**

Based on clustering result, we compared each attributes with the mean of the user groups and observes 4 different personalities:

A. The Play-safes:Group1 who tends to have fewer loans, mortgages than average. They tend to be younger.
B. The High-touchers: Group2 who inquiry more often than other groups. Its age distribution is heavy on millenials.
C. The High-flyers: Group3 which has higher credit limit and mortgages than any other groups. Group3 are also more mature in age and have higher credit scores.
D. The Econnoissuers: Group4 which tends to have multiple cards and products, presumably to compare and find the best deal


Cluster
1
2
3
4
count_bankruptcy
count_inquiries_3_months
count_inquiries_6_months
count_inquiries_12_months
count_open_installment_accounts_24_months
count_total_tradelines_opened_24_months
count_tradelines_cc_opened_24_months
count_tradelines_closed_accounts
count_tradelines_condition_derogatory
count_tradelines_open_collection_accounts
count_tradelines_open_mortgages
count_tradelines_open_secured_loans
count_tradelines_open_student_loans
count_tradelines_open_unsecured_loans
count_tradelines_opened_accounts
max_cc_limit
total_auto_loans_balance
total_cc_open_balance
total_mortgage_loans_amount
total_mortgage_loans_balance
total_open_cc_amount_past_due
total_student_loans_balance
total_tradelines_amount_past_due
total_tradelines_open_balance
tradelines_avg_days_since_opened

### Data ethics

We did not include sensitive information such as zipcode which could leads to discrimination. We looked at generic population-level demographics such as age bucket to better understand the needs of Credit Sesame customers.

### Business strategies

Knowing different user segments allows us to more accurately target each group's needs and improve their user experience, while at the same time feed the product offerer's income stream.

The Econnoissuers are the majority and have high propensity to accept new products, as such we can show them product comparisons to save their time in decision making, and present them with recommended better alternatives to build brand royalties with Credit Sesame.

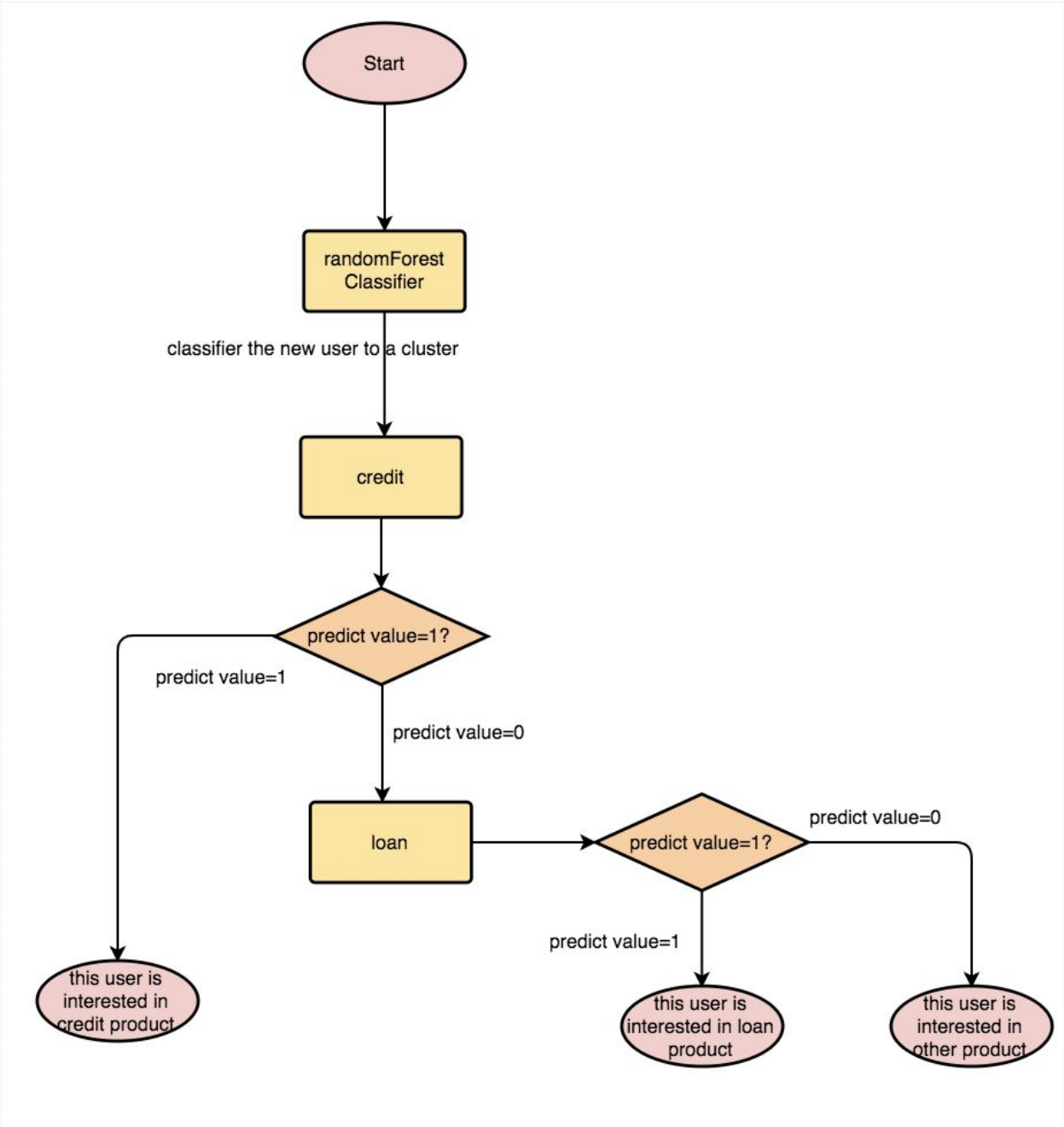
The High-flyers are seasoned financial product users and possibly busy professionals. For this group we need to target with useful information, fewer product and less frequent product, as the same time mitigate their risk as they tend to have higher loans.

The high-touchers have more questions than the other groups and we can make chat-bot or real-person chat support option more obvious to them to aid them with every step.

The play-safes are the minority and we can target them with moderate amount of offering and observe their reaction, then posteriorly adjust communication style with this groups of user. The posterior update applies to all groups.

## Predictions

In order to predict the signup by user segments and product, we looked into which users interact with which product offer page and built customized submodels to inspect what leads to the ultimate sign-up. Our modeling process can be seen from the image below, which can be mainly divided to 3 sub-models(random forest classifier, credit classifier and loan classifier).




## RandomForest Classifier

***Usage:** We want to use this classifier to classify the new user to the most similar cluster based on our clustering result.

***Method:** We build this model based on random forest algorithm and use the grid search method to adjust model's parameters.

***Code:**

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV

para ={'n_estimators':[50,100,150,200]}
gsearchRandom = GridSearchCV(estimator = RandomForestClassifier(random_state=42),param_grid = para,cv=5)
gsearchRandom.fit(x_train, y_train)
pred=gsearchRandom.predict(x_test)
```

## Credit Classifier

***Usage:** This model can be used to predict whether a user is interested in credit products or not, including credit cards and other credit products.

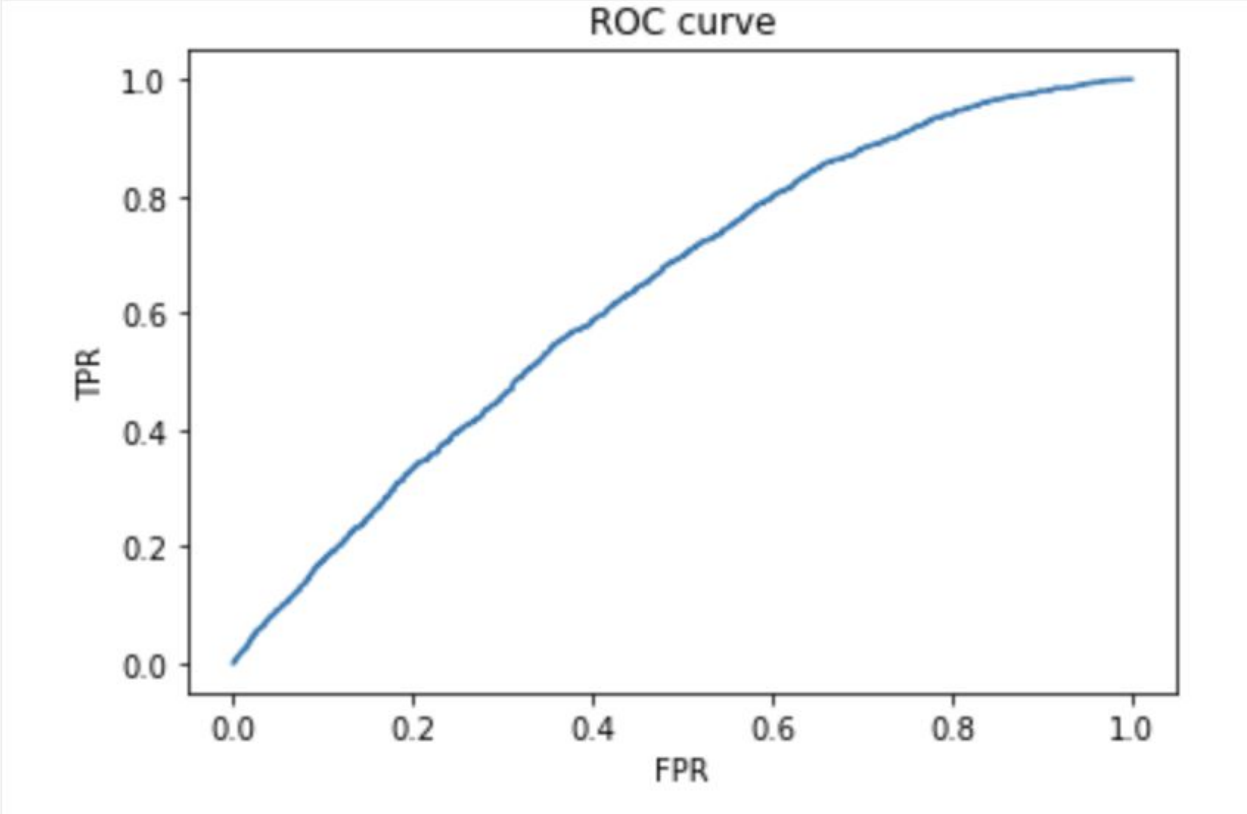
***Method:** We build this model based on Ridge classification algorithm and Gradient Boosting Classification algorithm and used grid search method to adjust model's parameters.

*** Performance**

*** Ridge Classification**

```
RidgeClassifier(alpha=0.5, class_weight=None, copy_X=True, fit_intercept=True,
        max_iter=None, normalize=False, random_state=None, solver='auto',
        tol=0.001)
```

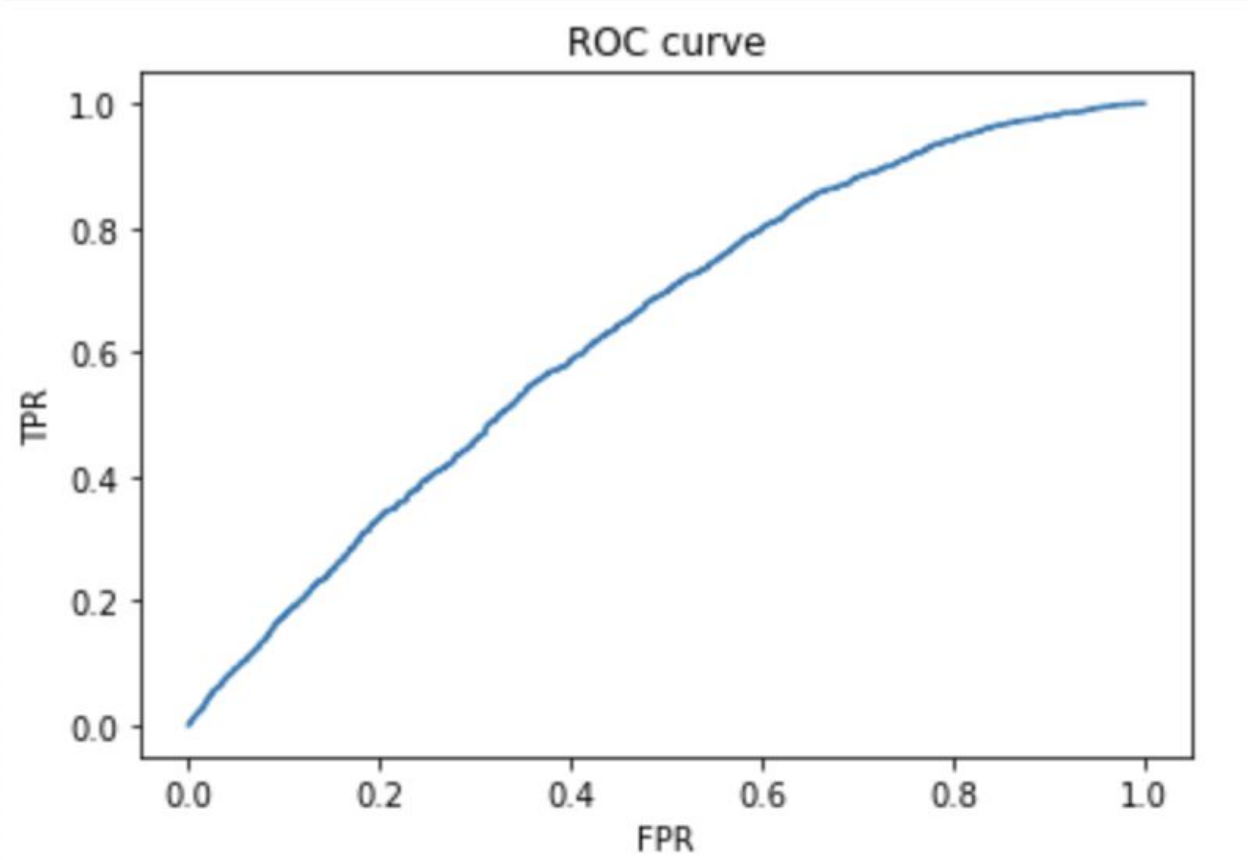
Test accuracy: 0.637
accuracy = 0.698786581014
sensitivity = 0.699141630901
specifity = 0.533333333333
PPV = 0.998569677156
NPV = 0.00378967314069
F1 score = 0.822450353416




* **Gradient Boosting Classification**

```
GradientBoostingClassifier(criterion='friedman_mse', init=None,
              learning_rate=0.1, loss='deviance', max_depth=8,
              max_features='sqrt', max_leaf_nodes=None,
              min_impurity_split=1e-07, min_samples_leaf=20,
              min_samples_split=300, min_weight_fraction_leaf=0.0,
              n_estimators=80, presort='auto', random_state=10,
              subsample=0.8, verbose=0, warm_start=False)
```

Test accuracy: 0.637
accuracy = 0.698786581014
sensitivity = 0.699141630901
specifity = 0.533333333333
PPV = 0.998569677156
NPV = 0.00378967314069
F1 score = 0.822450353416




## Loan Classifier

***Usage:** This model can be used to predict whether a user is interested in loan products or not, including personal loan, personal loans and other mortgage.

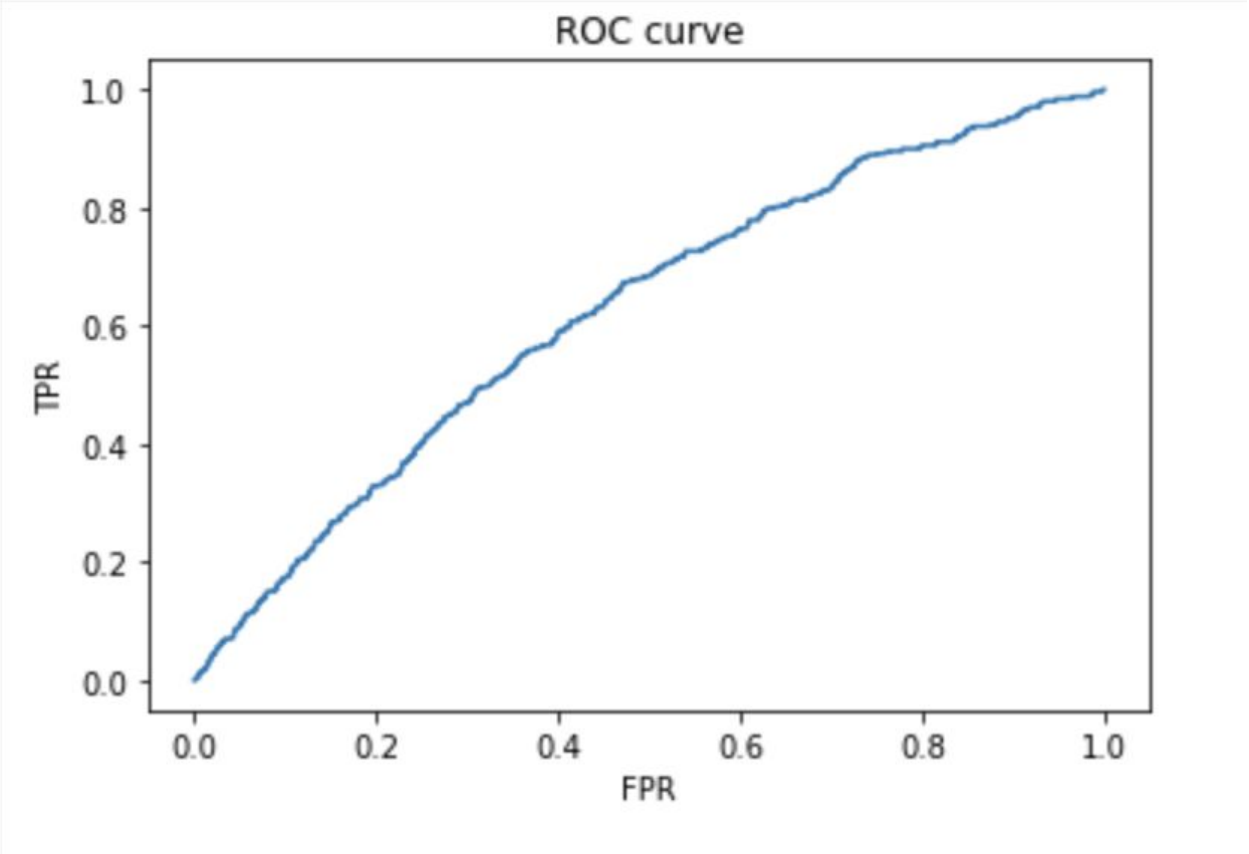
***Method:** Just like how we build credit model, we build this model based on Ridge classification algorithm and Gradient Boosting Classification algorithm and used grid search method to adjust model's parameters.

*** Performance**

*** Ridge Classification**

```
RidgeClassifier(alpha=0.05, class_weight=None, copy_X=True,
        fit_intercept=True, max_iter=None, normalize=False,
        random_state=None, solver='auto', tol=0.001)
```

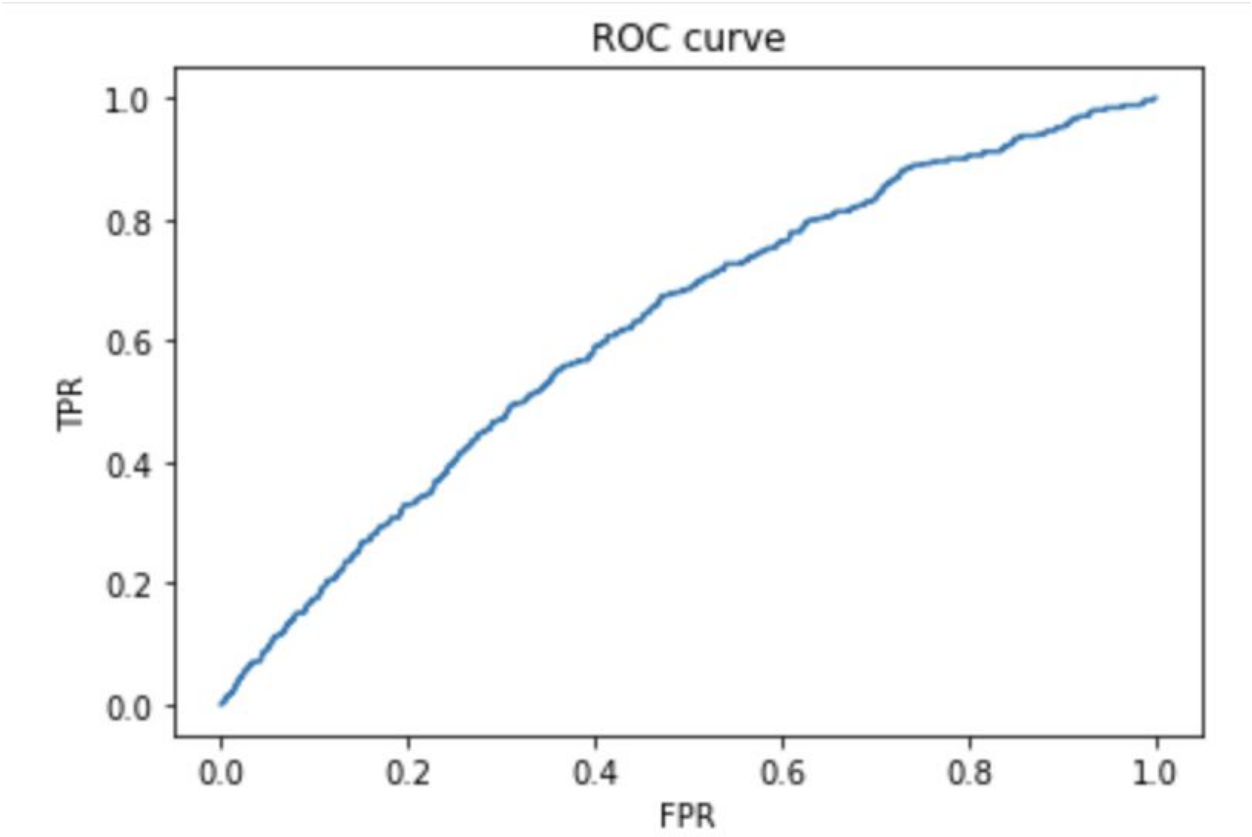
Test accuracy: 0.624
accuracy = 0.893621575342
sensitivity = 0.893621575342
specifity = nan
PPV = 1.0
NPV = 0.0
F1 score = 0.943822764779




* **Gradient Boosting Classification**

```
GradientBoostingClassifier(criterion='friedman_mse', init=None,
              learning_rate=0.1, loss='deviance', max_depth=8,
              max_features='sqrt', max_leaf_nodes=None,
              min_impurity_split=1e-07, min_samples_leaf=20,
              min_samples_split=300, min_weight_fraction_leaf=0.0,
              n_estimators=70, presort='auto', random_state=10,
              subsample=0.8, verbose=0, warm_start=False)
```

Test accuracy: 0.615
accuracy = 0.893621575342
sensitivity = 0.893621575342
specifity = nan
PPV = 1.0
NPV = 0.0
F1 score = 0.943822764779




### Business Implications

- We can build other models with the similar functions to predict whether the new user is interested in other products or not. Due to the time limit, we will not elaborate as the steps are similar
- The company can market to the target users with different interest. For example, if the predicting model shows one user may interested in credit product, the company can send more ads about credit products to this user.

## Next step

As a next step we can:

1. Build a real-time clickstream visualization system to monitor the evolution of user behaviors
2. Look into user bounce rate by page which could be achieve with a better understanding of how current webpages are organized.
3. Since the performance of our models still have potential to improvement, we can conduct further parameter tuning.

## Tools:

R, Python, Tableau, Arcgis