Miguel Granica

Capstone Proposal

Starbucks Capstone Challenge

Domain Background

- Within the business strategy of the Starbucks company, we focus on the permeability of its offers for a limited group. The datasets provide us with information on both the demographic characteristics of consumers and their receptivity to different types of offers.
- Within the supervised models of Machine learning, these types of problems refer to Classification issues. It consists of a predictive model that infers a target class from a data set
- This model is highly applied in all kinds of disciplines.
    - Email spam detector
    - Conversion prediction
    - Movie review classification
    - MRI images classification
    - e-commerce, customer reviews analysis…
- We refer to an example that describes a case study of opinion polarity classification:
    - https://www.researchgate.net/publication/328306943_A_Comparison_of_Machine_Learning_Algorithms_in_Opinion_Polarity_Classification_of_Customer_Reviews

Problem Statement

- Using the information provided in the datasets, we intend to infer which way a specific client will respond to a certain type of offer.
- On one hand, We are able to establish a demographic segmentation based on the different categorical fields such as age, gender, income
- On the other, we can link these demographic segments to the type of offer and their associated receptivity

Datasets & Inputs

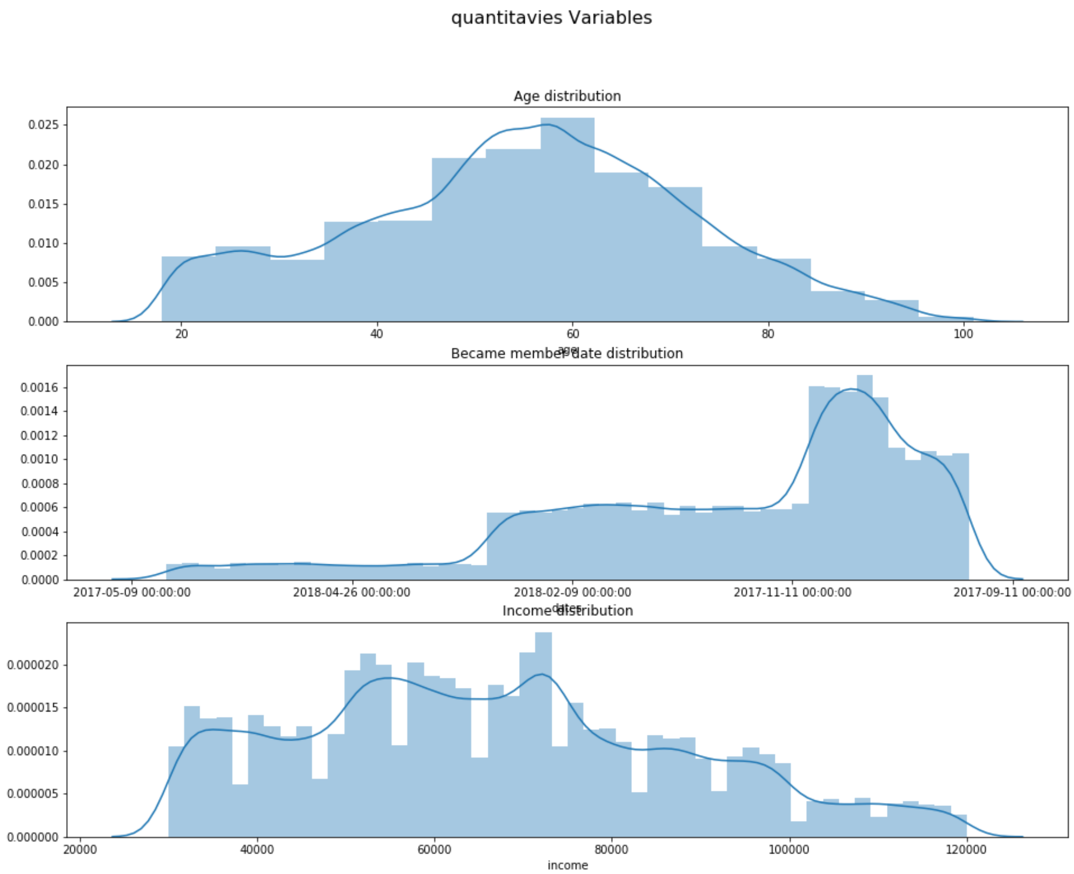
The data is contained in three files:

- **portfolio.json -** containing offer ids and meta data about each offer (duration, type, etc.)
    - *id* (string) - offer id
    - *offer_type* (string) - type of offer ie BOGO, discount, informational
    - *difficulty* (int) - minimum required spend to complete an offer
    - *reward* (int) - reward given for completing an offer
    - *duration* (int) - time for offer to be open, in days
    - *channels* (list of strings)

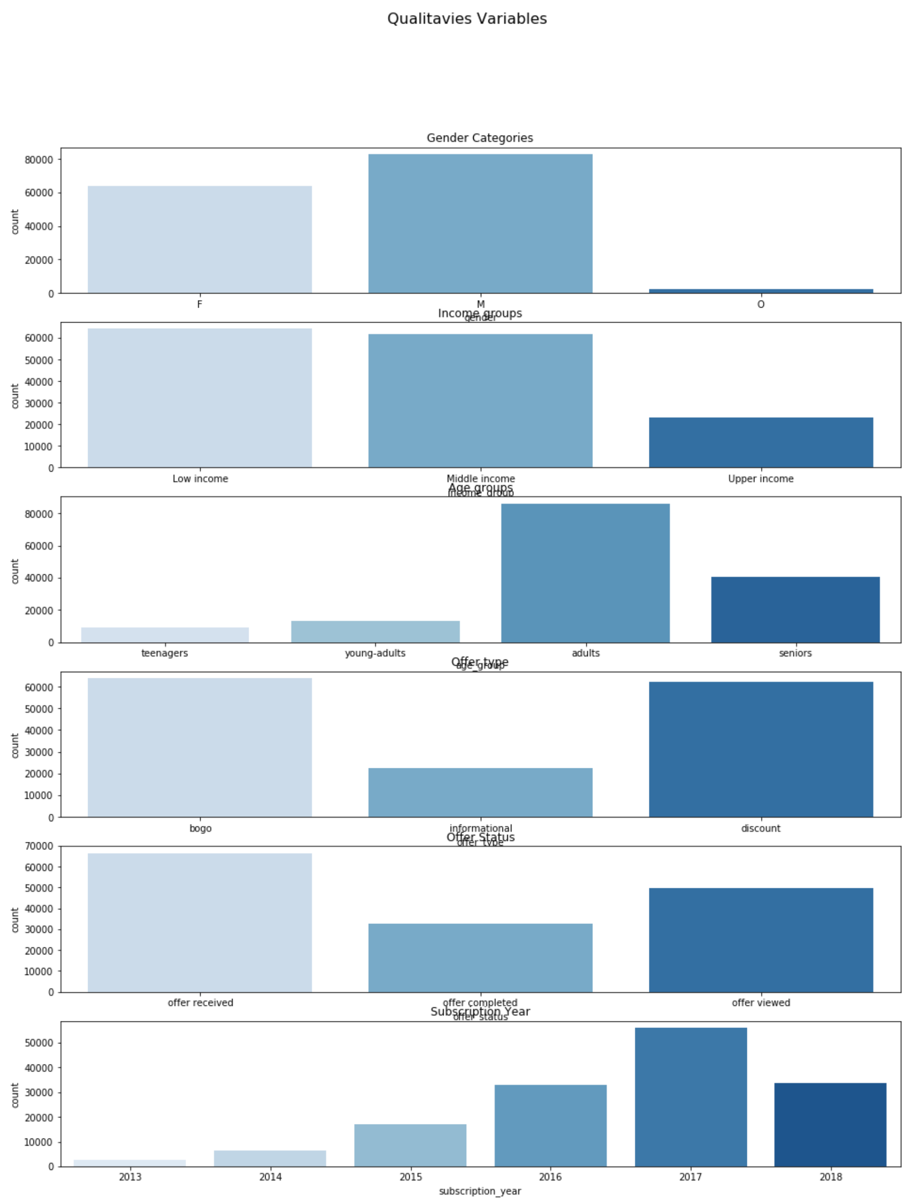
- **profile.json** - demographic data for each customer
    - *age* (int) - age of the customer
    - *became_member_on* (int) - date when customer created an app account
    - *gender* (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
    - *id* (str) - customer id
    - *income* (float) - customer's income

- **transcript.json** - records for transactions, offers received, offers viewed, and offers completed
    - *event* (str) - record description (ie transaction, offer received, offer viewed, etc.)
    - *person* (str) - customer id
    - *time* (int) - time in hours since start of test. The data begins at time t=0
    - *value* - (dict of strings) - either an offer id or transaction amount depending on the record

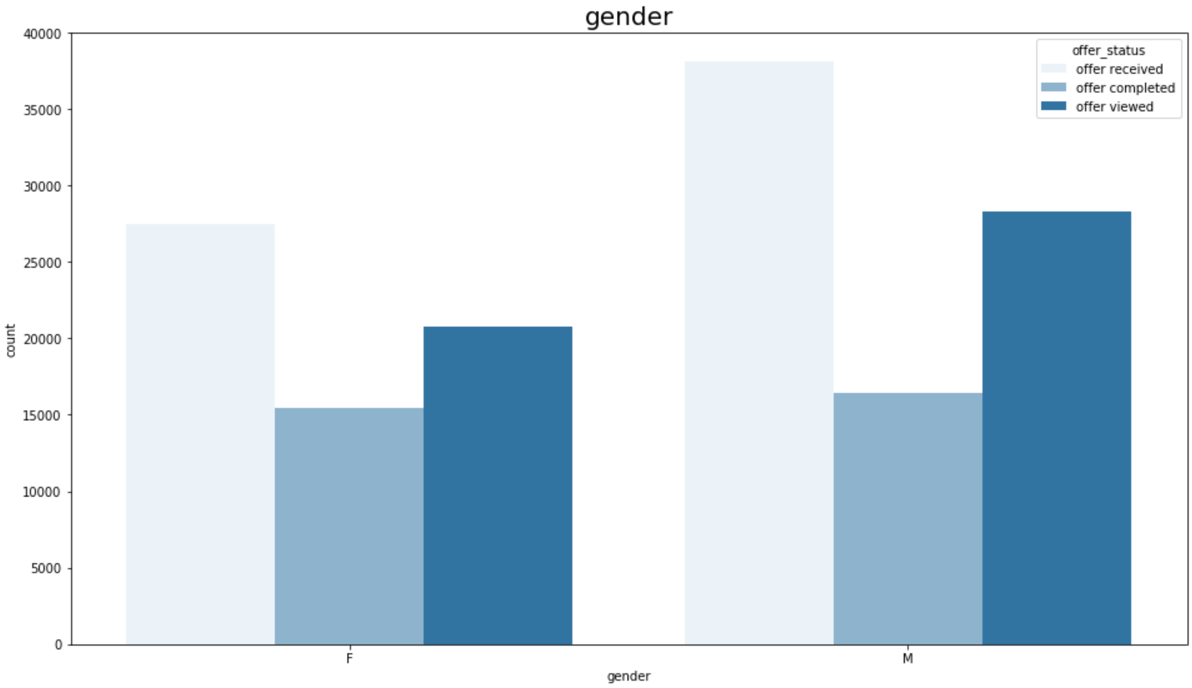
- In a first approximation, we focus on obtaining the distributions of the numerical variables:
    - age tends towards a symmetrical distribution with its center around 60 years and with a large sigma
    - the distribution of subscription dates is right skewed, it has its peak at the end of 2018
    - Income is distributed multimodally

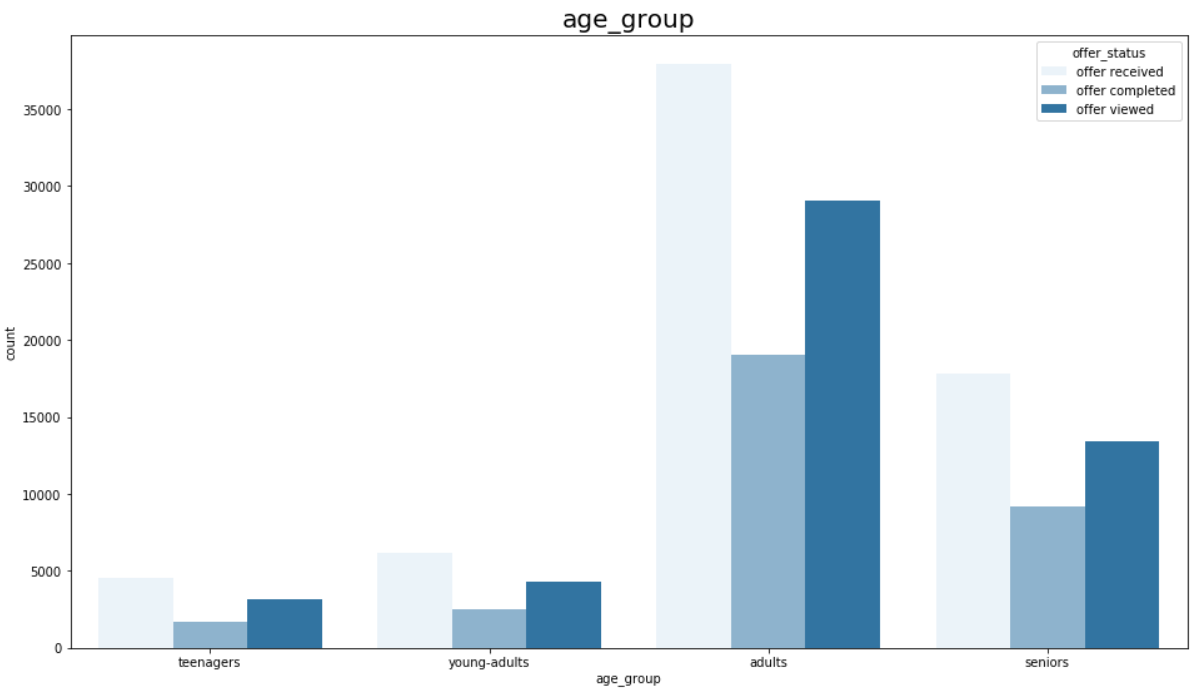


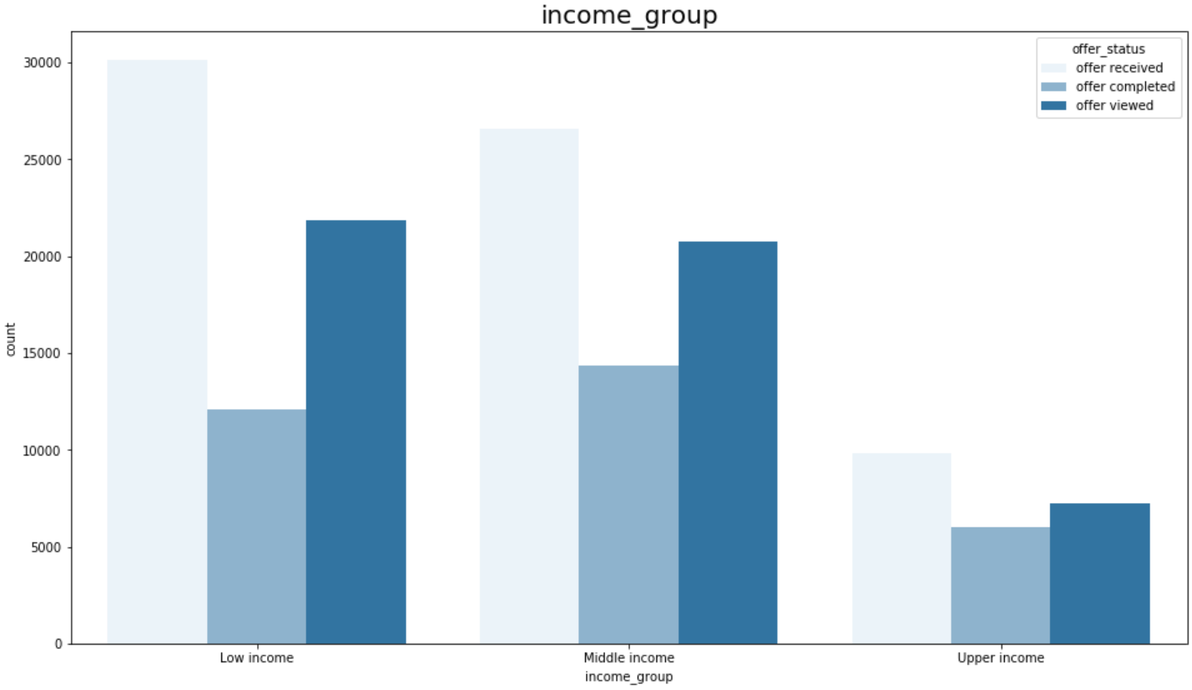
- secondly, and to build a general perspective of the features, we analyze the categorical variables:
    - Men with 80,000 interventions are represented 20% more than women. around 2000 cases do not have a defined gender
    - The age group that will consent to the greatest number of cases is that of adults with more than the friendship of the participations. for population segmentation analyzes this bias should be considered
    - with respect to income, upper income has 50% fewer interventions than the other 2
    - the type of offer is divided unevenly: 40 BOGO 40 discount 20 informational
    - the conversion ratio seems less than 50% but we are going to develop it and segment it throughout this notebook
    - There has definitely been a growth in the number of subscribers to strabucks between 2013 and 2018, peaking in 2017

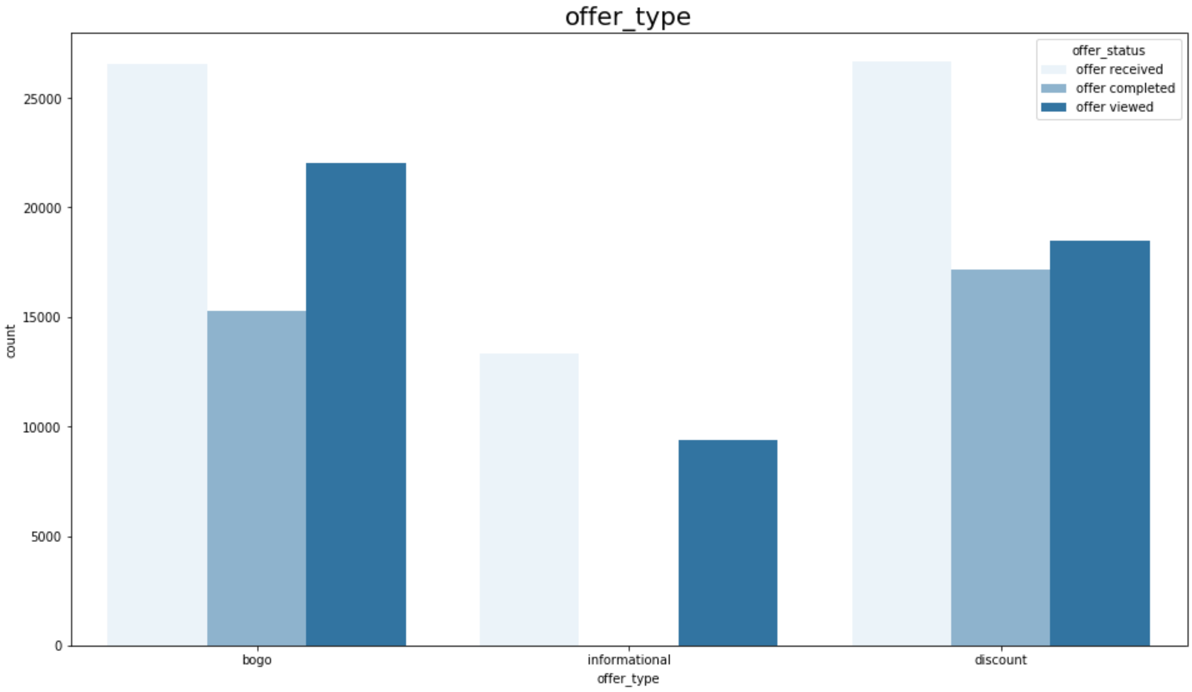


- At this point we focus on the effectiveness and type of offers based on some demographic aspects:
  - On average, men receive 25% more offers than women. However, the ratio between offers seen and completed among women is lower. which shows a better reception of them for the female gender
  - Adults and seniors are the age groups that accumulate more than 95% of the cases: 70% and 25% respectively. However, the age group with the best response to offers is the senior
  - The type of offer that has the best impact on customers seems to be the discount






income_group
offer_status
offer received
offer completed
offer viewed
count
30000
25000
20000
15000
10000
5000
0
Low income
Middle income
Upper income
income_group


offer_type
offer_status
offer received
offer completed
offer viewed
count
25000
20000
15000
10000
5000
0
bogo
informational
discount
offer_type

- Summary table: demographic segmentation of offer status

  We decided to build an interactive dataframe using ipywidgets to have a better understanding of the behaviors that different consumer groups have. For this, we are first going to group the dataframe by the categorical fields that identify demographic groups:

  - gender
  - age
  - income

  and by the fields that describe the characteristics of the established offers:

  - offer type
  - offer status

			offer_status	offer completed	offer received	offer viewed	offer conversion
gender	age_group	income_group	offer_type				
F	teenagers	Low income	bogo	191.0	357.0	295.0	53.50
			discount	225.0	341.0	213.0	65.98
			informational	NaN	189.0	125.0	NaN
		Middle income	bogo	104.0	144.0	119.0	72.22
			discount	100.0	149.0	89.0	67.11
			informational	NaN	61.0	37.0	NaN
	young-adults	Low income	bogo	345.0	546.0	446.0	63.19
			discount	334.0	511.0	319.0	65.36
			informational	NaN	293.0	178.0	NaN
		Middle income	bogo	147.0	222.0	178.0	66.22
			discount	137.0	209.0	130.0	65.55
			informational	NaN	103.0	70.0	NaN
	adults	Low income	bogo	1238.0	2027.0	1724.0	61.08
			discount	1373.0	2074.0	1433.0	66.20
			informational	NaN	1018.0	757.0	NaN
		Middle income	bogo	2009.0	2782.0	2366.0	72.21
			discount	2195.0	2879.0	2191.0	76.24
			informational	NaN	1427.0	1069.0	NaN
		Upper income	bogo	1096.0	1412.0	1129.0	77.62
			discount	1094.0	1379.0	966.0	79.33
			informational	NaN	729.0	488.0	NaN

Solution Statement

- The objective will be to build a machine learning model that allows identifying, based on the different demographic segments, the result of the offers offered to customers.
- To model the predictions about this problem we required supervised Machine learning algorithms. we will use classification algorithms
    - Logistic regression
    - Support Vector Machine
    - K-Nearest Neighbors
    - Decision Tree
    - random forest

Benchmark Model

- In order to obtain target labels the algorithm set as a benchmark model is a naive model

Evaluation metrics

	Decision Tree	**Random Forest**	**Logistic Regression**	**Support Vector Machine**	**Naive Bayes**	**K-Nearest Neighbors**
Accuracy	100	100	100	86.7	100	83.8
Precision	100	100	100	91.6	100	83.1
Recall	100	100	100	72.9	100	73.7
F-Measure	100	100	100	81.2	100	78.1

Project design

- Data Cleaning
    - portfolio dataframe:
        - Rename id column name to offert_id and set as index
        - Hot-encoding the offer_type column
        - Hot-encoding the channels column
    - profile dataframe:
        - Drop age values == 118
        - Drop NaN values for income and gender columns
        - Dateformat became_member_on
        - Binary values for gender column
    - transcript dataframe:
        - unstack amount and offer id columns
        - set time unit in days

- Master table consolidation

	gender	age	income	subscription_year	offer_id	time	amount	reward	difficulty	duration	email	mobile	social	web	offer_type
1	1	55.0	112000.0	2017	1	22	0.0	5.0	5	7	1	1	0	1	1
3	1	75.0	100000.0	2017	1	0	0.0	0.0	5	7	1	1	0	1	1
4	1	75.0	100000.0	2017	1	5	0.0	5.0	5	7	1	1	0	1	1
6	2	68.0	70000.0	2018	1	17	0.0	0.0	5	7	1	1	0	1	1
7	2	68.0	70000.0	2018	1	21	0.0	5.0	5	7	1	1	0	1	1

- Exploratoriy Data Analysis
    - Univariable exploration
    - Bivariable exploration
    - Segmentation demographics and offer status summary table

- Machine learning preprocessing
    - Based on the different categorical fields, how the different demographic segmentations respond to each one of the types of offers present in the exercise, we can build a Machine learning model that indicates how a certain demographic profile would respond to the different types of offers
    - To validate the impact that Starbucks promotions have, we are going to exclude received offers from the study, in this way we will only take into account the offers seen over the completed ones.
- ML train&fit the model
    - Even if we decide to do a binary classification we could also have tried to do Multiclass classification for the offerts status:
        - Logistic regression
        - Support Vector Machine
        - K-Nearest Neighbors
        - Decision Tree
        - random forest
    - by the evaluation metrics it looks like we have overfitted the model, an useful tool to apply for feature reduction would be the PCA, and with this implementation apply some hyperparameter tuning to sharpen the model