Technical blog for project of hackathon:

Credit card consumption prediction for 3rd quartile:

By Team Cyberbots:

* Ruchita Patil

Education: Electronics and Telecommunication Engineering from Mumbai University(2015)

Work Experience: CRM Consultant@ Zycus(2015-present)

* Heera Kesavan

Education: Electronics and Telecommunication Engineering from Mumbai University(2016)

Work Experience: Call coordinater@ Softcell Technologies(2017-2018),

Technical Consultant office 365@Wipro(2018-present)

* Komal Bharadva

Education: Electronics Engineering from Mumbai University(2016)

Work Experience:Business Dev. Consultant @Robokart.com(jun2016-feb-2017)

Associate Consultant@ Capgemini (March-2017 to present)

* Ankita Dutta Gupta

Education: Electrical & Electronics Engineering from Visvesvaraya Tech. University(fresher-2020)

**Problem Statement:**

We have data from Bank with customer details, like age, gender and other demographics. Also, information on liabilities, assets and history of transactions with the bank for each customer.

We have to Predict the averagespend for different set of customers for next 3 months(July, August & September), to understand the relationship between the customer profile and their spending patternsin order to gain insights and generate decisions based on the customer persona and their Expenditure patterns.

About the dataset:

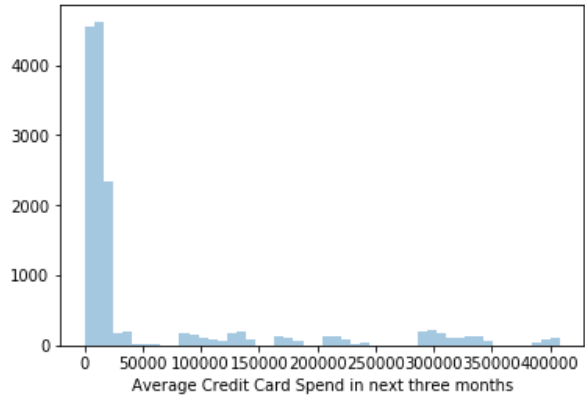
The data consists of records of 15000 clients and 44 features. There are 43 predictors and 1 target that describes expect the average credit card expenditure of the client for the 3rd quarterly.



**Approach towards the dataset:**

**EDA:**

* **Target Variable :**

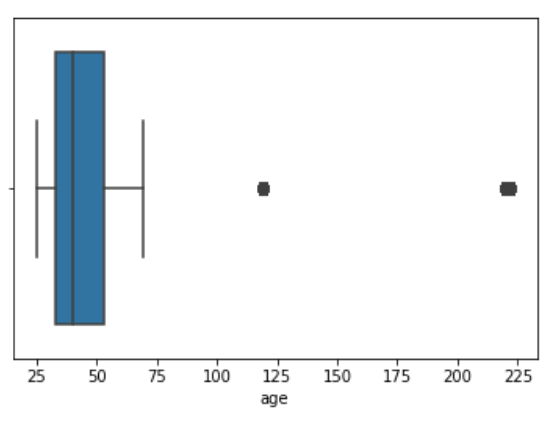


(distplot of target variable)  
  
From the available data ,we have the expected amount of expenditure of our client base for the next 3 months. Approx amount of 50k spent by majority of population over 3 months.The higher and lower values of the range are :

4,08,382 and 5,341 respectively.

* **AGE:**

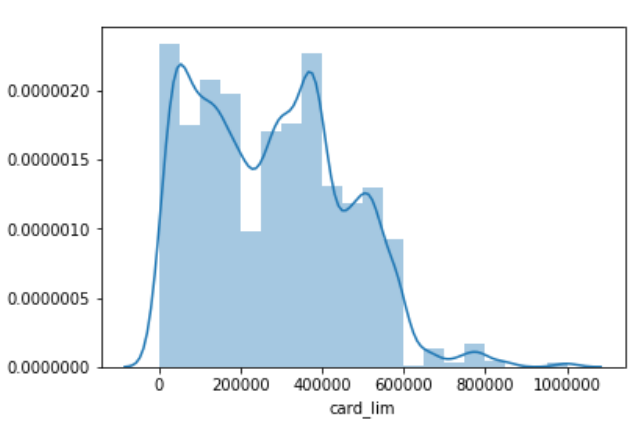
1. About the age: Age range of the client dataset ranges from 25 to 223 years
2. Maximum ages are between 25 to 55
3. Outliers are 125 and 223 for a very less percentage of population.
4. Feature moderation: dropped age greater than 120 since only 5 entries.
5. For ages greater than 70 we have replaced with the median value of the column which is 38 years



(boxplot of the age)

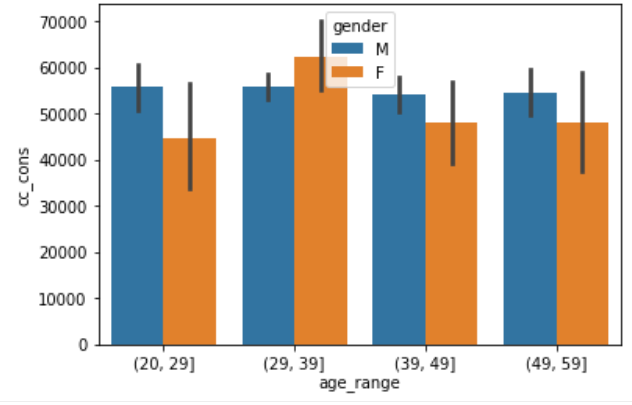
* **Card Limit**

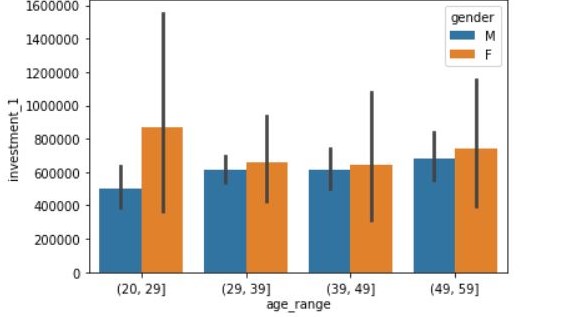
Details about the credit card limit set for the customers:  
   
Minimum limit: Rs. 10,000  
Maximum limit :Rs. 10 L  
Maximum population have limit between a range of   
10k to7 Lakhs  
Feature treatment : Removed zero values from the column.



(histplot of card limit)

* **Correlation of Age with the expected CC amount spend for the clients:**



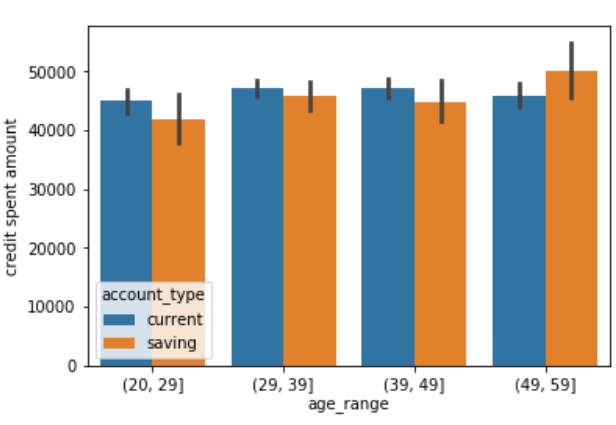


Observation:

* Females majorly in their 30s :expected to spend most (maybe because of financial stability?

Or highly lucrative CC offers on various retail fronts?)

* Males don’t show much deviation for the same.
* Females in their 20s have made maximum investment in their DEMAT account so they are likely to spend less in the subsequent months.
* **To study the trend for the client based on their type of account**



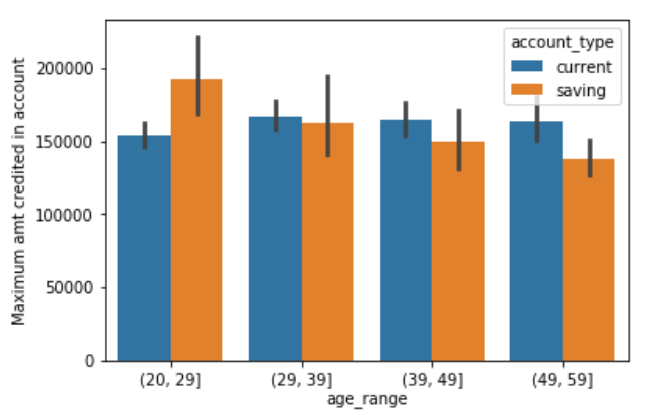
* Observation:
* Clients (in their 50s and having a savings type a/c) have spent the most using their CCs in

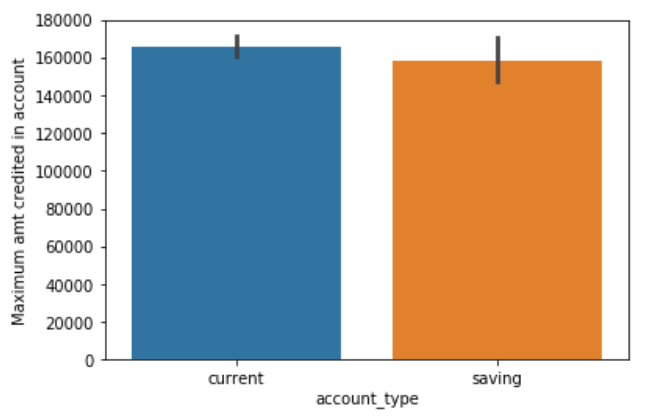
the past 3 months.

* This shows expenditure increases with the age for people having savings a/c(salaried

Employees/having monthly income)

* Traders/Entrepreneurs(who need to access their a/c Daily) having current a/c don’t show much deviation. (*Can we lower the fees and interests charged on the overdraft facilty To motivate current a/c Holders to spend more* ?)
* **Age with max credit card spent and type of account**





Steep decrease observed in the credited amount to a/c of clients with ‘savings’ type with increase in age range. Minimum deviation for the same for ‘current’ type a/c.  
Also there is overral increase in the amount credited for current a/c .

* **Credit card transaction exploration with respect to Region to find if any trend exists**:

1. Region Code 424 has maximum amount of transaction of 11,61,694.11 with # of transaction 95( Avg ~ 12,228)

2. Region Code 424 has minimum amount of transaction of 1544 with # of transaction 192( Avg ~ 8)

3. Region Code 505 has minimum number of transactions using credit which is 3 , amount spent is 1,01,083( Avg ~ 33,694)

4. Region Code 713 has maximum number of transactions using credit card which is 1220 , amount spent is 66,735 (Avg ~ 55)

5. Clearly average of region 505 > average of region 424

which makes us largely presume that region 505 is most affluent of all.

**Baseline Model:**

As the Target Variable was continuous, we used Linear Regression for our Baseline Model. We obtained this model without performing any feature engineering.

As our Evaluation criteria was RMSLE (Root Mean Squared Logarithmic Error) Score,after building the baseline model with Linear Regression we obtained the score as 7.44.

**Feature Engineering:**

**1)**There were no missing values in our Dataset.

**2)**We dropped the following columns as they had only 1 unique value and they had least contribution to predict our Target Variable:

*personal\_loan\_active,vehicle\_loan\_active,personal\_loan\_closed,vehicle\_loan\_closed,loan\_enq*

**3)**We found the correlation of all the features present in our dataset with Target variable.

Region code is least correlated and cc\_cons\_may is most correlated with the target.

**4)** We calculated interquartile range (IQR), score to **filter out the outliers** by keeping only valid values.

The interquartile range (IQR), also called the midspread or middle 50%, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 − Q1.

**5)**We have taken the average of amount spent by Credit card and Debit card respectively in past 3 months i.e. April, May, June as below:

**df['cc\_cons\_avg']=(df["cc\_cons\_jun"]+df["cc\_cons\_apr"]+df["cc\_cons\_may"])/3**

**df['dc\_cons\_avg']=(df["dc\_cons\_jun"]+df["dc\_cons\_apr"]+df["dc\_cons\_may"])/3**

**6)**Since the Column “investment\_4” has negative values we have taken the absolute of this column.

**Evaluation Criteria: RMSLE (Root Mean Squared Logarithmic Error) Score**

The RMLSE isthe relative Error between the predicted and the actual values.

**Model Selection**

After outlier treatment and feature selection, we tried various models and obtained the following results:

|  |  |
| --- | --- |
| **Modelling method** | **RMSLE score** |
| Linear regression | 2.05 |
| Logistic regression | 2.18 |
| Random Forrest Classifier | 2.92 |
| Linear SVC | 2.1 |
| AdaBoostRegressor | 2.39 |

We chose Linear Regression as our final modelling algorithm, as it gave us the best RMSLE Score.

* **Understanding after problem solving:**
* To understand the credit card expenditure for a general age group of people
* Which type of bank account is more likely to see more credit card expenditure
* Types of investments made by the people and the trend in their respective expenditures.
* Also which age group needs to be targeted by the bank for the future offers.

**Insights and decisions by our team:**

* Customers to be targeted: Age Range:30-50 years
* Motivate clients of 30s age group to open DEMAT using their saving a/c
* Can Pitch high end offers to the affluent regions
* Provide in depth analysis and stock recommendation from expert analysts.

LinkedIn url of team:

Ankita: <https://www.linkedin.com/in/ankita-duttagupta-46a730169>

Ruchita: <https://www.linkedin.com/in/ruchita-patil-91111810b/>

Komal: <http://linkedin.com/in/komal-bharadva-99b8b6106>

Heera: [https://www.linkedin.com/in/heera-kesavan-100329129/](https://slack-redir.net/link?url=https%3A%2F%2Fwww.linkedin.com%2Fin%2Fheera-kesavan-100329129%2F)