BFSI Capstone Project

Final- Submission

Submitted By: Imran Khan

Business understanding

- CredX is a leading credit card provider that gets thousands of credit card applications every year. But in the past few years, company has experienced an increase in credit loss.
- Objective is to 'acquire the right customers' in order to decrease credit loss to the company.
- Hence, we need to identify the customers who are very less likely to Default on their Credit Card payments with the help of Predictive Modeling there by determining the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of your project.

Approach

High-level

- Based on the Data provided and the problem statement outlined, it's a binary classification problem.
- We are given two datasets – Demographic and Credit Bureau both have same number of Records with unique Applicant's ID.
- Analyze each datasets one after another but before we merge the clean datasets together, we plan on building few Models on Demographic data to access the predictive powers of the variables using **Logistic, SGD-Classifier, Decision Forest, Random Forest, XG Boost, Cat Boost, AdaBoost and Light GBM classifiers.** We may decide to stop at first 2-3 algorithms mentioned above if we found that the dataset is not performing well for the business objective we are trying to solve.
- We will initially try to build the Model and access performance using standard process of NULL values imputations with the help of KNN (K-Nearest Neighbors), Dummy variables creating, Standardization to see if the approach gives us good results in the beginning.
- We also need to build Models and access performance by replacing actual values with the corresponding WOE values calculated for each variable with respect to the Target Variable which in our case is "Performance Tag".
- Upon thoroughly inspecting the Demographic Dataset in terms of Outcome from EDA and various performance metrics obtained from different algorithms, we will move onto merge Demographic dataset with Credit Bureau Data to build final models.
- We followed **CRISP- DM (CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING)** framework to accomplish all the above mentioned approach.
    - Business Understanding
    - Data Understanding
    - Data Preparation
    - Data Modelling
    - Model Evaluation
    - Model Deployment (This depends on business if they want to take the model forward in production. Hence, this will not be covered.)

Data Understanding

1) <u>As mentioned before, we are given 2 datasets to accomplish the objective:</u>
   - Demographic Data: This contains information provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status and other profile related details.

   - Credit Bureau Data: This contains information from the credit bureau and contains variables such as customers who were 30/60/90 days past delinquent, Credit card utilizations, Outstanding Balance, presence of different types of loans...etc.

Below is the snapshot of the Data Dictionary:

Credit Bureau Data:

Credit Bureau Data	
Variable	**Description**
Application ID	Customer application ID
No of times 90 DPD or worse in last 6 months	Number of times customer has not payed dues since 90days in last 6 months
No of times 60 DPD or worse in last 6 months	Number of times customer has not payed dues since 60 days last 6 months
No of times 30 DPD or worse in last 6 months	Number of times customer has not payed dues since 30 days days last 6 months
No of times 90 DPD or worse in last 12 months	Number of times customer has not payed dues since 90 days days last 12 months
No of times 60 DPD or worse in last 12 months	Number of times customer has not payed dues since 60 days days last 12 months
No of times 30 DPD or worse in last 12 months	Number of times customer has not payed dues since 30 days days last 12 months
Avgas CC Utilization in last 12 months	Average utilization of credit card by customer
No of trades opened in last 6 months	Number of times the customer has done the trades in last 6 months
No of trades opened in last 12 months	Number of times the customer has done the trades in last 12 months
No of PL trades opened in last 6 months	No of PL trades in last 6 month of customer
No of PL trades opened in last 12 months	No of PL trades in last 12 month of customer
auto loans)	Number of times the customers has inquired in last 6 months
auto loans)	Number of times the customers has inquired in last 12 months
Presence of open home loan	Is the customer has home loan (1 represents "Yes")
Outstanding Balance	Outstanding balance of customer
Total No of Trades	Number of times the customer has done total trades
Presence of open auto loan	Is the customer has auto loan (1 represents "Yes")
Performance Tag	Status of customer performance (" 1 represents "Default")

Demographic Data:

Demographic Data	
Variables	**Description**
Application ID	Unique ID of the customers
Age	Age of customer
Gender	Gender of customer
Marital Status	Marital status of customer (at the time of application)
No of dependents	No. of childrens of customers
Income	Income of customers
Education	Education of customers
Profession	Profession of customers
Type of residence	Type of residence of customers
No of months in current residence	No of months in current residence of customers
No of months in current company	No of months in current company of customers
Performance Tag	Status of customer performance (" 1 represents "Default")

2) Data Sanity and Quality Checks:

- **Demographic Data:**
    - ➔ A total of 71295 rows and 12 columns.
    - ➔ 3 Applicant's IDs are duplicated. This was treated by taking the row that had most recent information with the help of Age column.
    - ➔ We have around 2% of the data missing for the Target variable and the remaining ones are shown in below screenshot.

```
Application ID                                   0.00
Age                                              0.00
Gender                                           0.00
Marital Status (at the time of application)      0.01
No of dependents                                 0.00
Income                                           0.00
Education                                        0.17
Profession                                       0.02
Type of residence                                0.01
No of months in current residence                0.00
No of months in current company                  0.00
Performance Tag                                  2.00
```

    - ➔ Age variable, has wrong data as it doesn't correlates well with the Education column. A person who has done PHD can't be of 15 years. All Age values less than 19 years are treated with the appropriate minimum age of the corresponding degree held by the applicant. There are few negative values as well in the Age column. Above approach takes care of that as well.
    - ➔ Income field has rows less than 0. These are treated by making them 4.5(minimum value in the distribution).
- **Credit Bureau Data:**
    - ➔ A total of 71295 rows and 19 columns.
    - ➔ 3 Applicant's IDs are duplicated. We retained only those rows based on the logic implemented for Demographic Data.
    - ➔ We have again around 2% of the data missing for the Target variable and the remaining ones are shown in below screenshot.

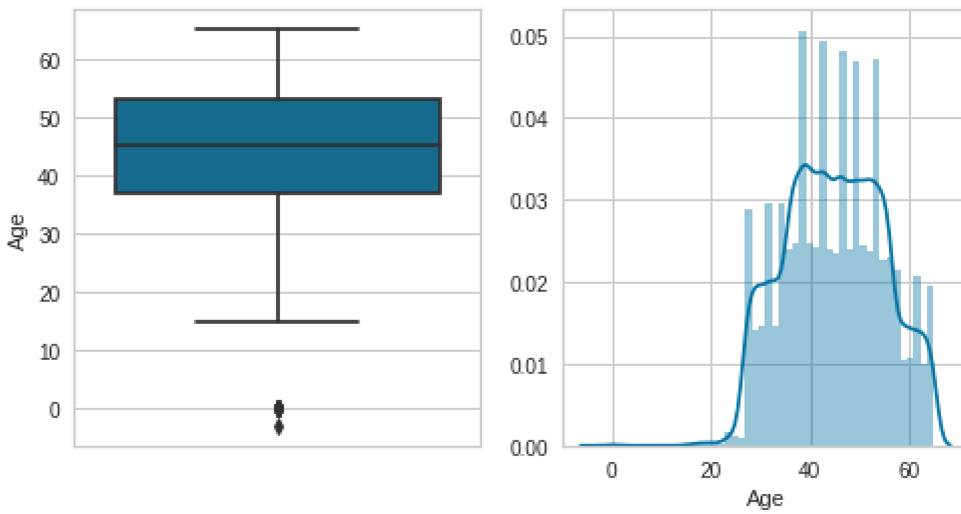
```
Application ID                                                  0.00
No of times 90 DPD or worse in last 6 months                    0.00
No of times 60 DPD or worse in last 6 months                    0.00
No of times 30 DPD or worse in last 6 months                    0.00
No of times 90 DPD or worse in last 12 months                   0.00
No of times 60 DPD or worse in last 12 months                   0.00
No of times 30 DPD or worse in last 12 months                   0.00
Avgas CC Utilization in last 12 months                          1.48
No of trades opened in last 6 months                            0.00
No of trades opened in last 12 months                           0.00
No of PL trades opened in last 6 months                         0.00
No of PL trades opened in last 12 months                        0.00
No of Inquiries in last 6 months (excluding home & auto loans)  0.00
No of Inquiries in last 12 months (excluding home & auto loans) 0.00
Presence of open home loan                                      0.38
Outstanding Balance                                             0.38
Total No of Trades                                              0.00
Presence of open auto loan                                      0.00
Performance Tag                                                 2.00
```

➔ Mostly numeric data. No Categorical variables in the dataset.

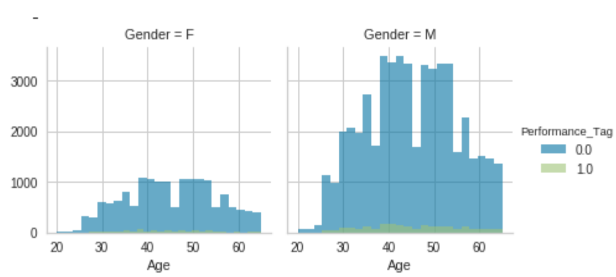
- In both the above two datasets we have removed the 2% performance Tag Null rows, considering these applicants were all denied the card. These data will be evaluated for each models build on the training sets. We can measure if our model is assigning Default tag to these applicants.
- For few we have replaced the data with the mode of the corresponding variables. And for some we have replaced the data with the WOE values of the most matching segment in the column.

3) <u>Exploratory Data Analysis -- **DEMOGRAPHIC** Datset:</u>

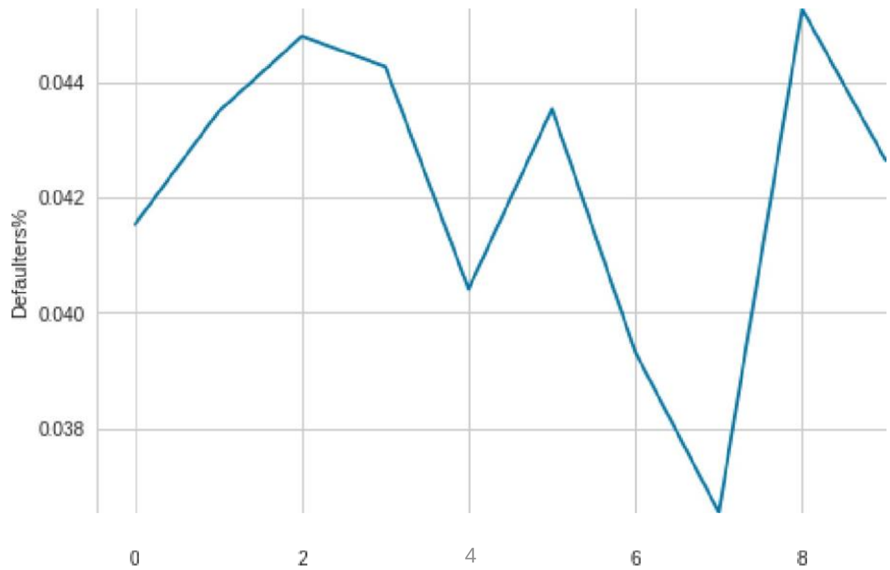
   - **<u>AGE</u>**:



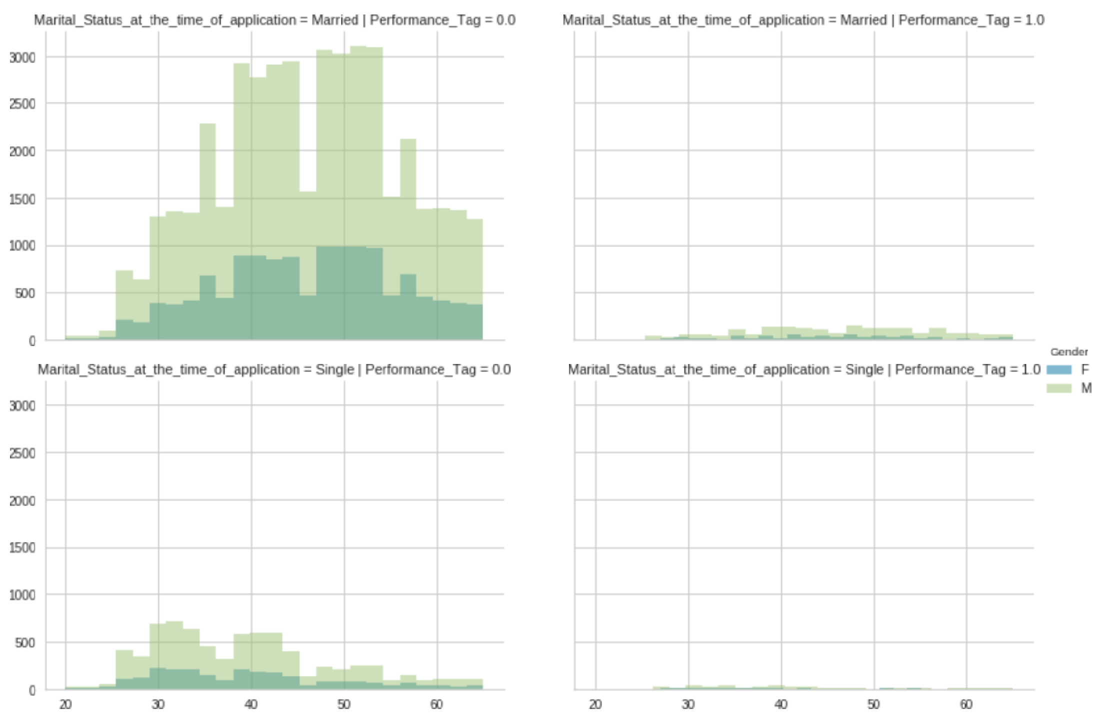
We can see that there are few outliers and the data also goes in negative.



- Based on above plot Population of Male applicant is substantially higher than Female.
- The Default Tag(Green color Bars) seems to be Higher for Male than Female. However, lets do some more analysis to look at actual numbers.

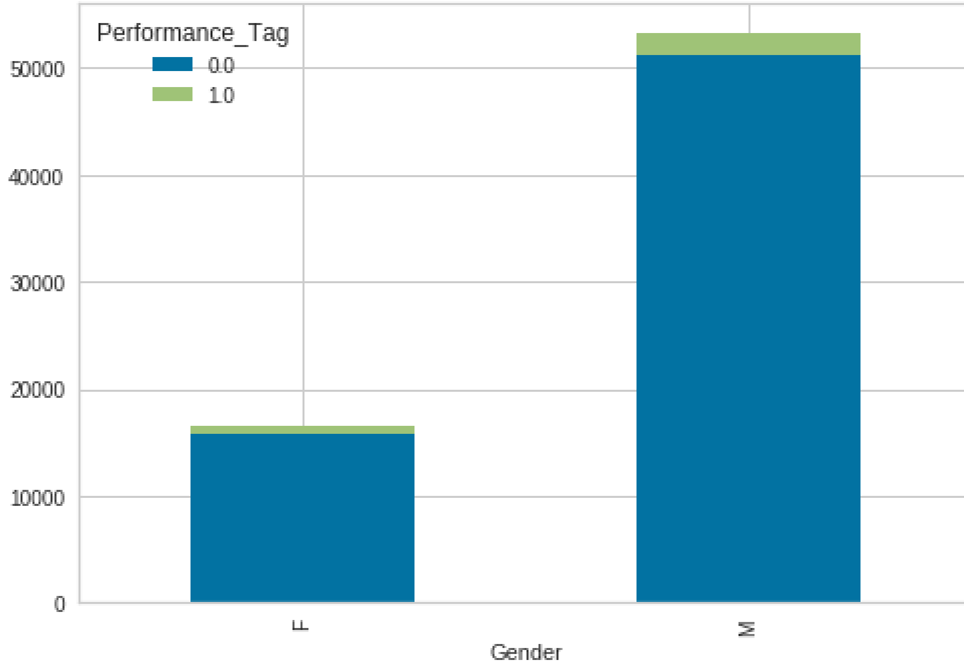


- Looks like the Defaulter °â is higher within either very young age group or very elderly people.



- We can see very clearly among MarFied people, Male tend to be majority class for both the Defaulters and Non-Defaulters Category.
- Also in Married category Male tend to be defaulting more than Female.
- We can see very dearly among Single people, Male tend to be the majority dass for both the Defaulters and Non-Defaulters Category who are defaulting more. This could be also due to the fact that the data is imbalanced.

➢ **<u>Gender</u>:**

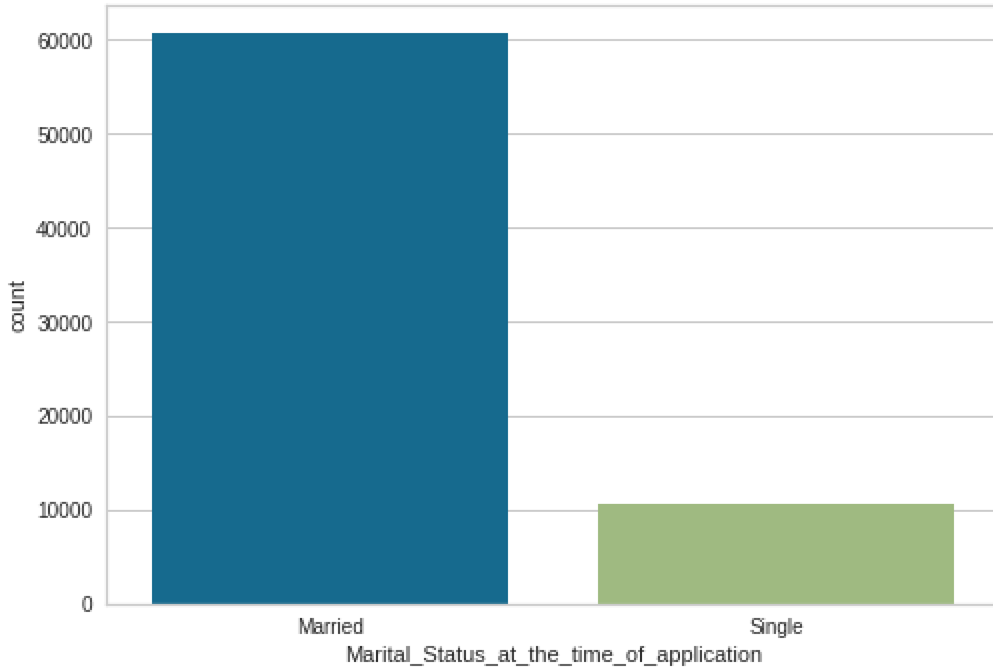


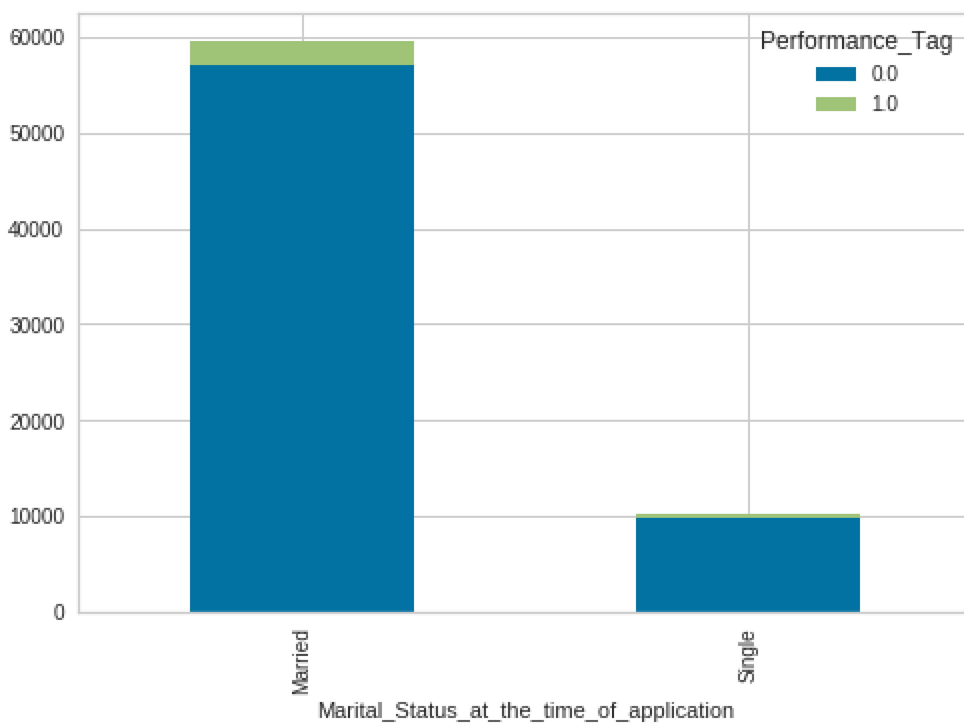
- Male Applicants are considerably higher than Female Applicants

Performance_Tag	0.0	1.0	perc
Gender			
F	15788	718	0.043499
M	51129	2230	0.041792

- Female seems to be Defaulting more but that could be due to less data points as well, Hence nothing conclusive can be inferred from this.

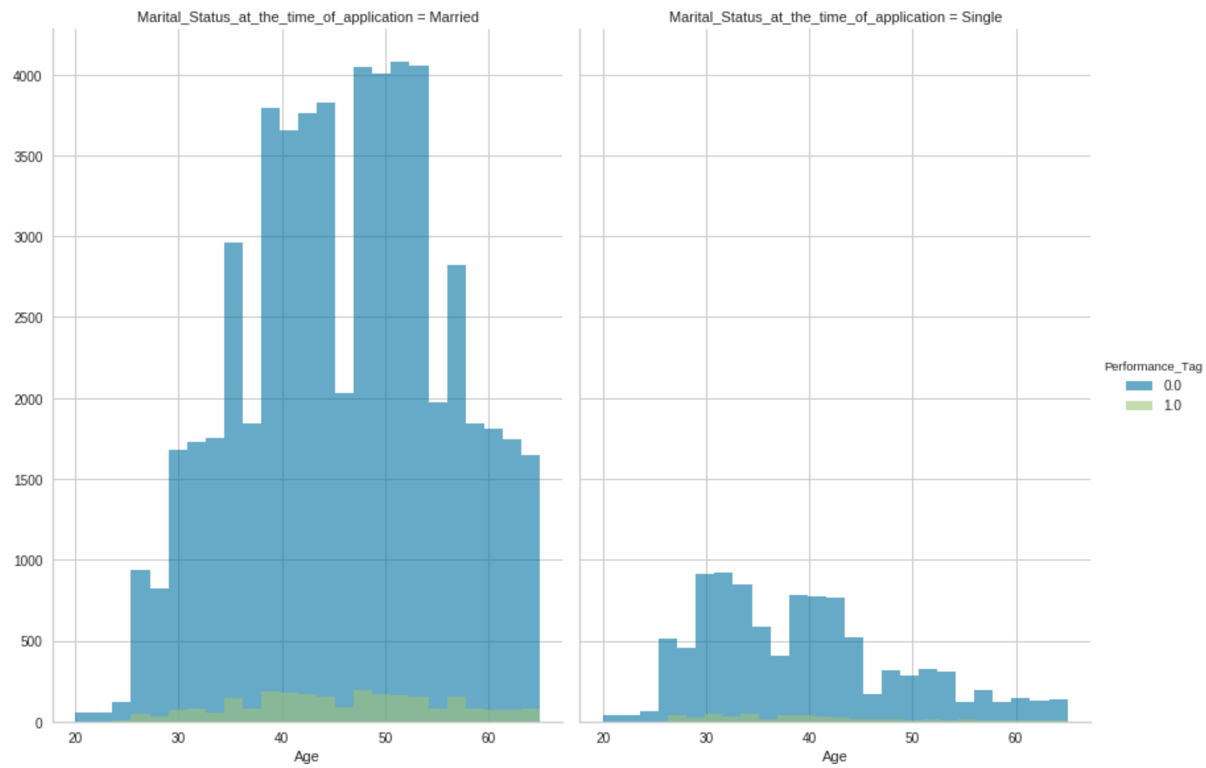
➢ **Marital_Status_at_the_time_of_application:**





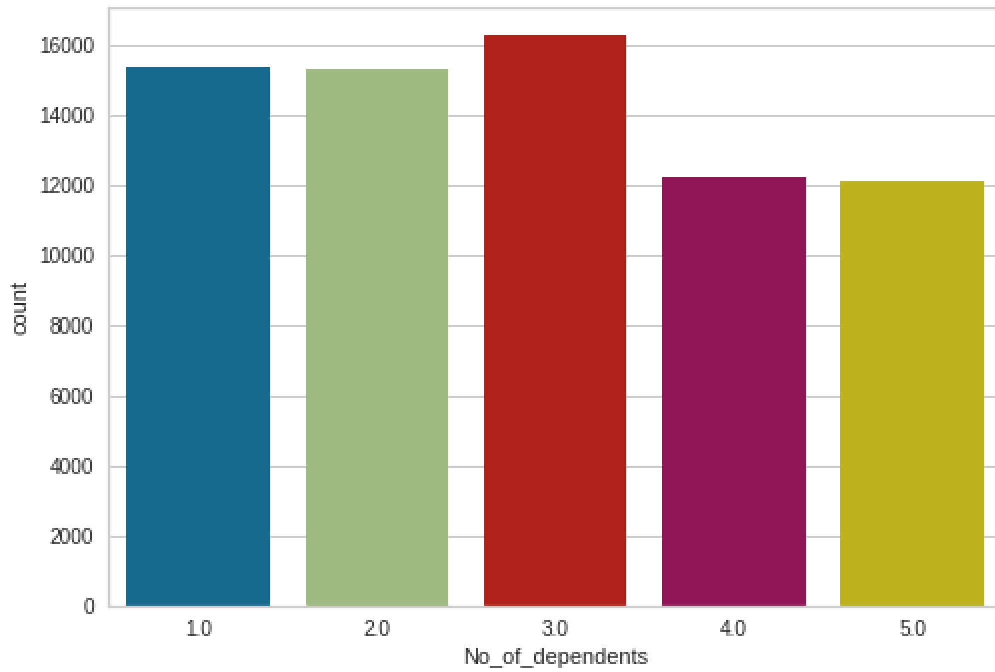
Performance_Tag	0.0	1.0	perc
Marital_Status_at_the_time_of_application			
Married	57041	2503	0.042036
Single	9872	445	0.043133

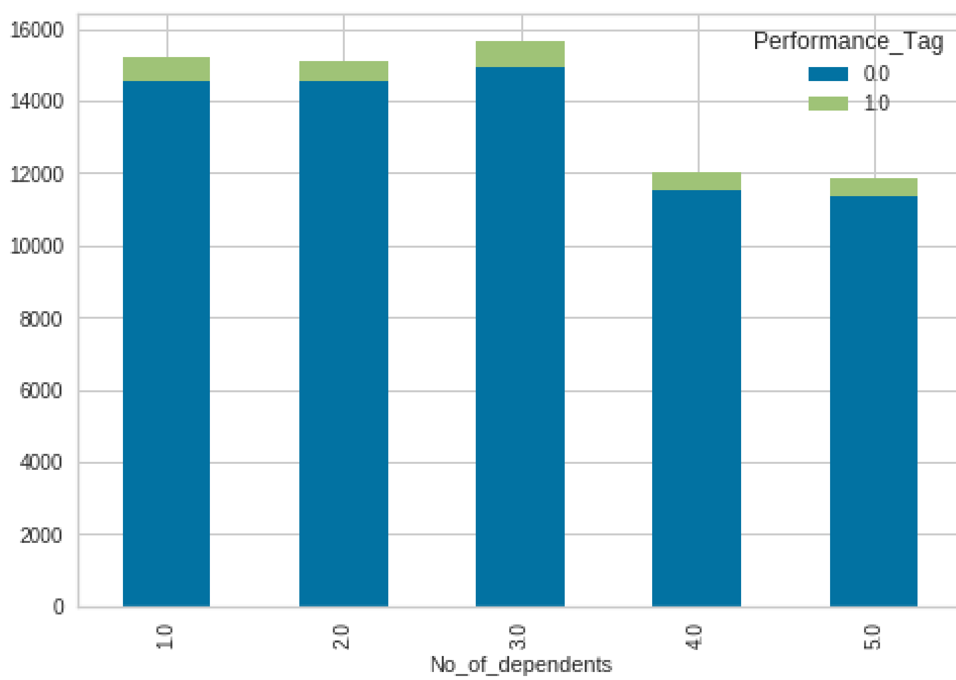
- Married Applicants are more than Singles.
- Single tend to default more compared to Married but as the count is not balanced, nothing conclusive can be inferred



Above plot shows the relationship between Age and Married/Single applicants with respect to the Defaulters and Non-Defaulters.

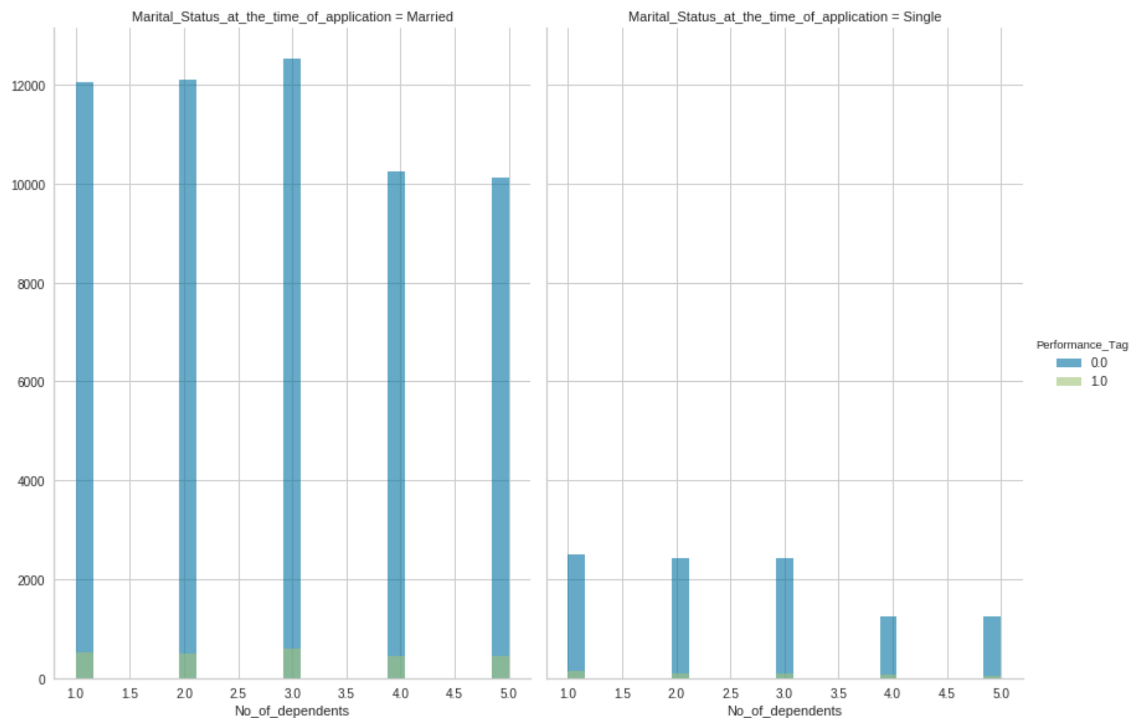
- **No_of_dependents:**



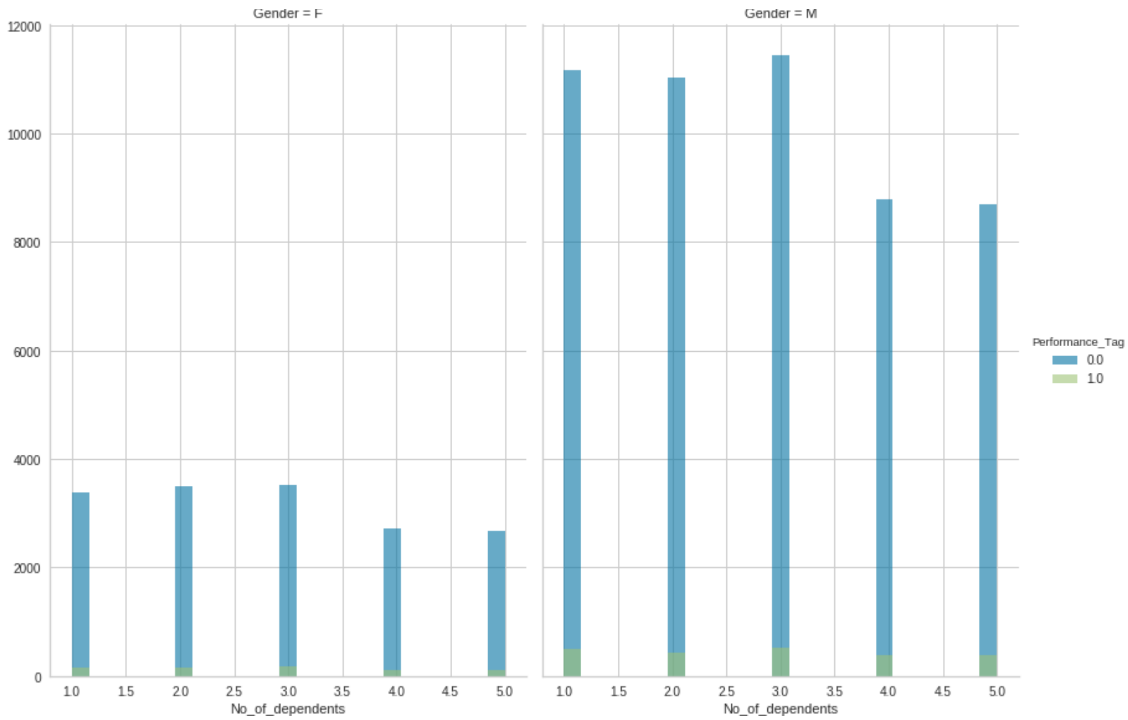


Performance_Tag	0.0	1.0	perc
No_of_dependents			
1.0	14551	667	0.043830
2.0	14539	588	0.038871
3.0	14949	695	0.044426
4.0	11505	494	0.041170
5.0	11372	504	0.042439

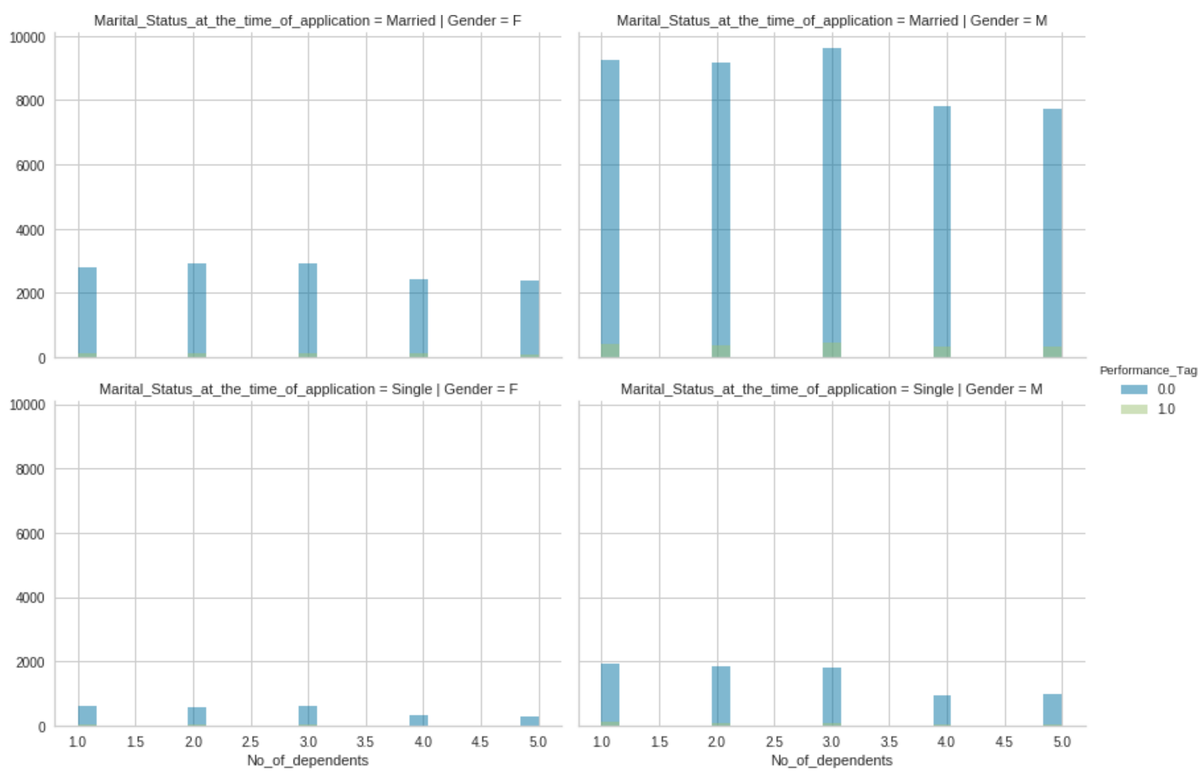
- Applicants with more dependents tend to default less as per the above stats but in Bivariate analysis we could come across some other findings about this pattern.
- Applicants with 3 dependents comparatively default more than rest of the category in thios segment.



Above plot shows the relationship between No. of Dependents, Marital status and Performance Tag

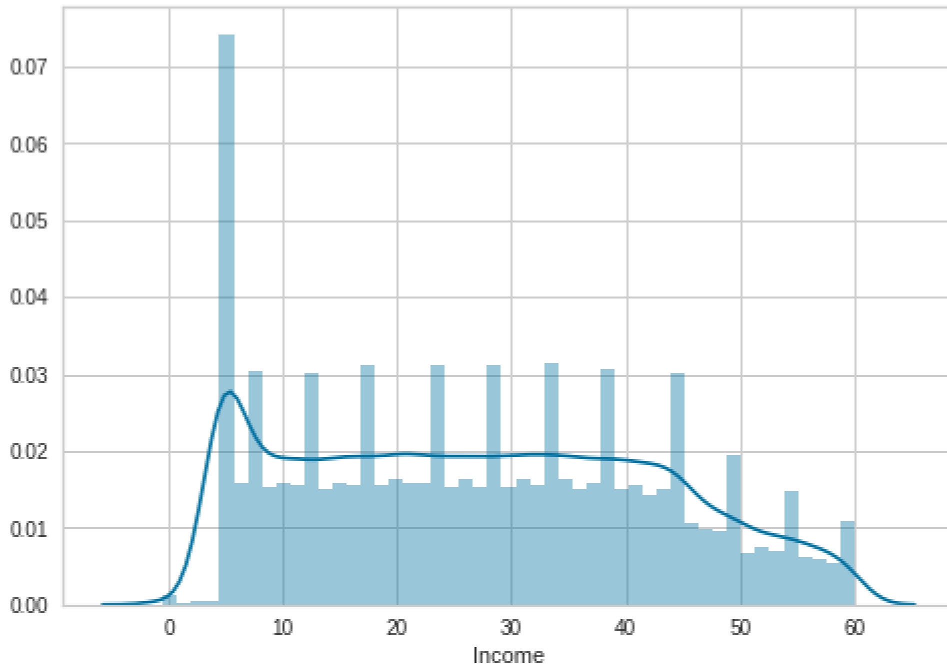


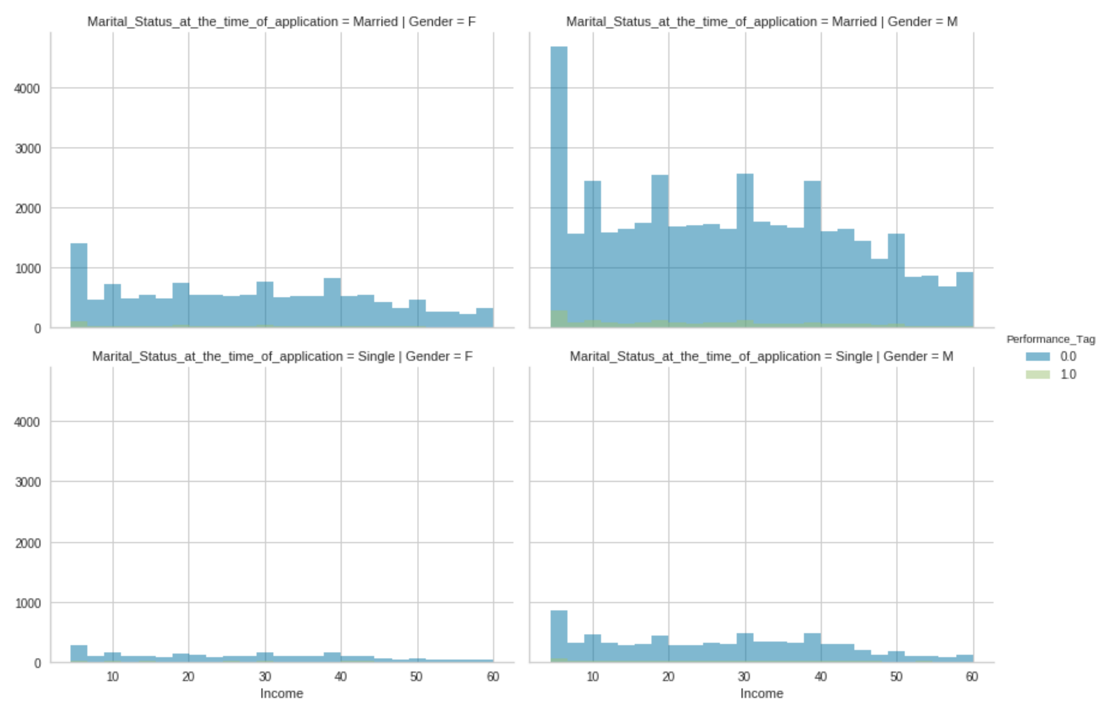
Above plot shows the relationship between No of Dependents, Gender and Performance Tag



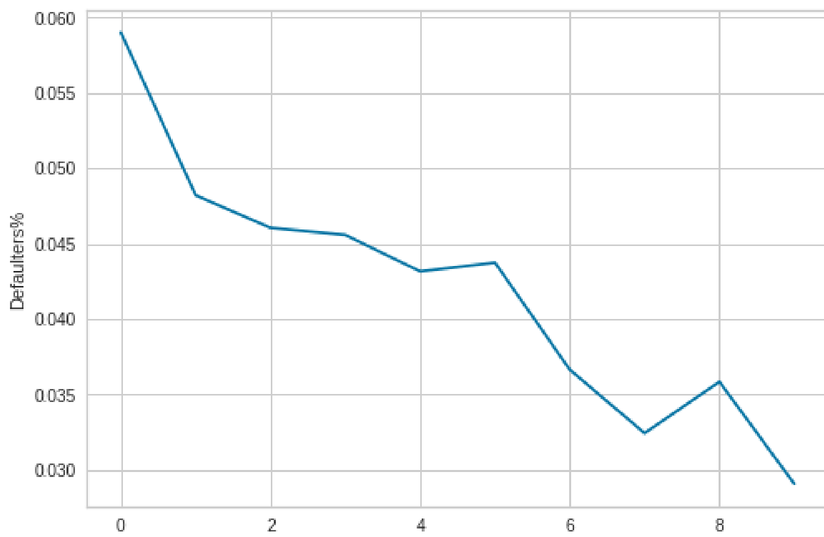
- **Male** and **Married** are the two categories which has maximum Default Rate

➢ Income:



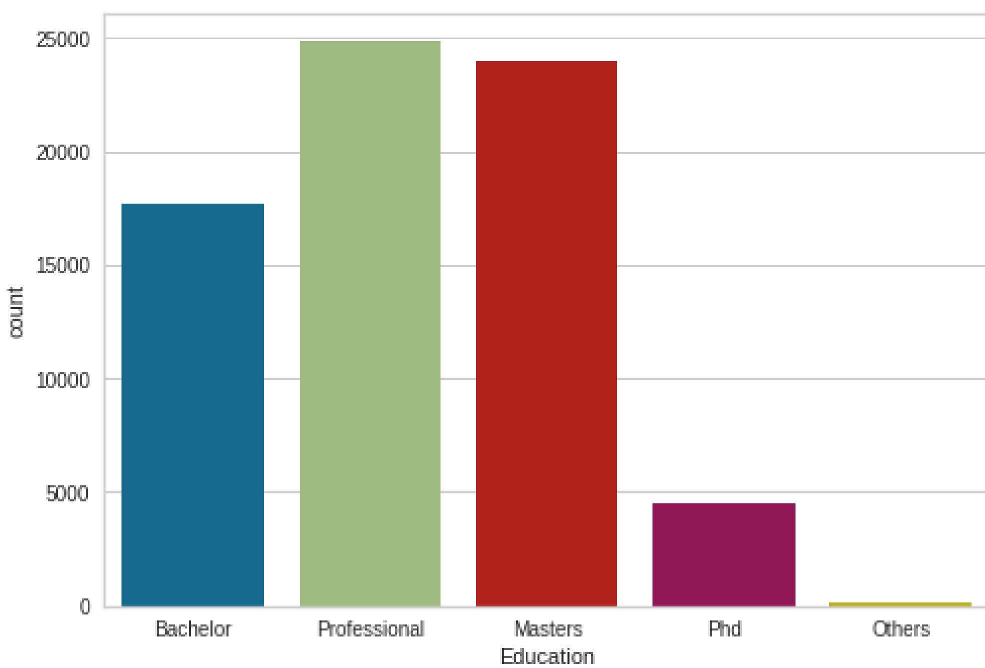


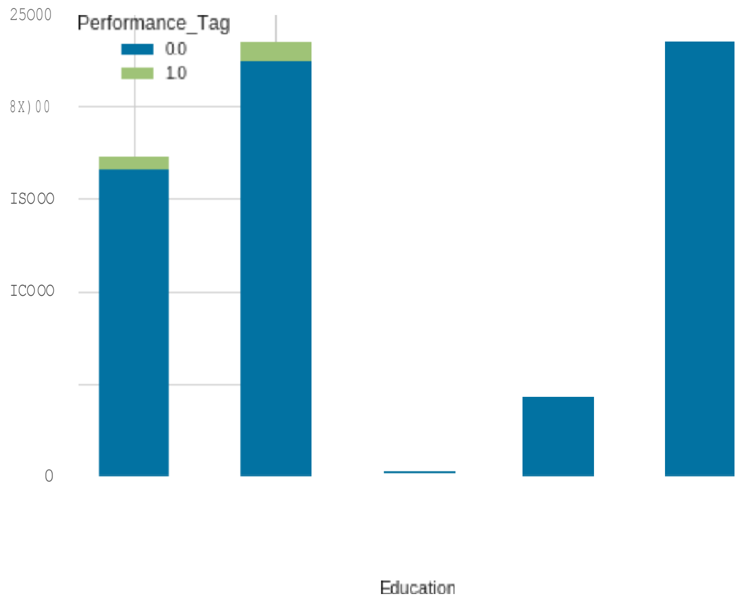
- We can see very clearly among **Married** people, Male tend to be majority class for both the Defaulters and Non-Defaulters Category.
- Also in **Married** category Male tend to be defaulting more in **Married** segment. Even in **Singles** Male Default more than Female.



- Looks like as the income increases, Default % also reduces.

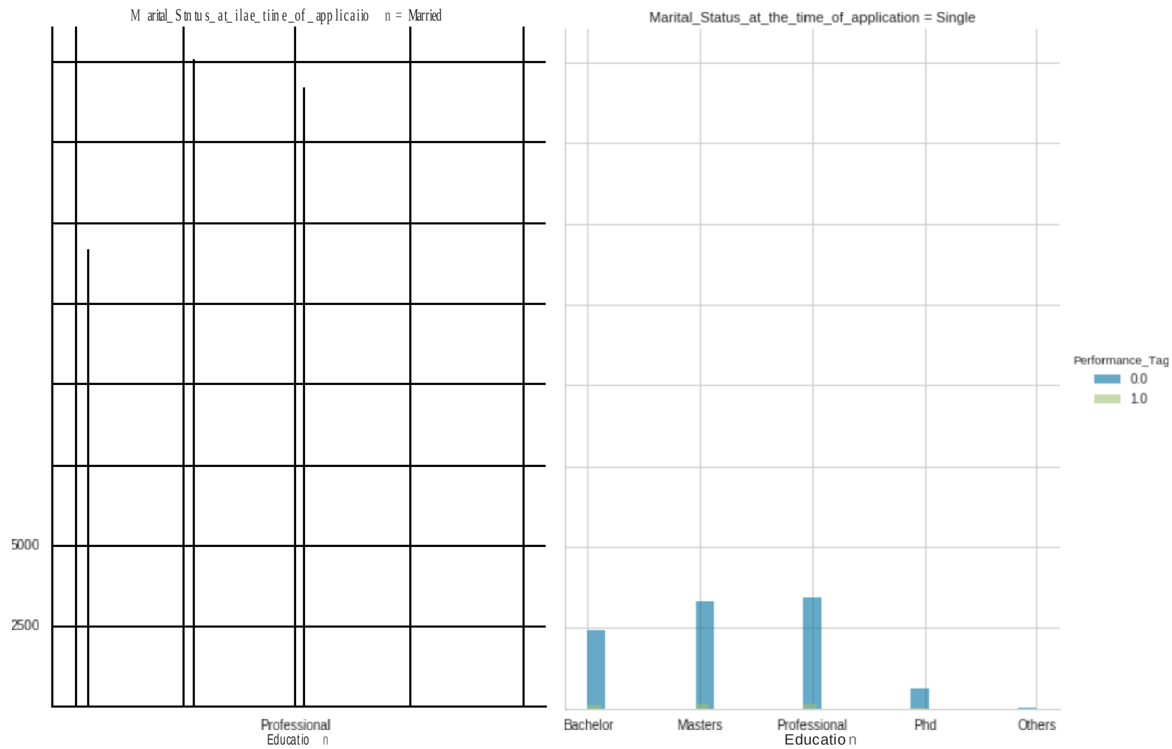
➢ <u>**Education:**</u>



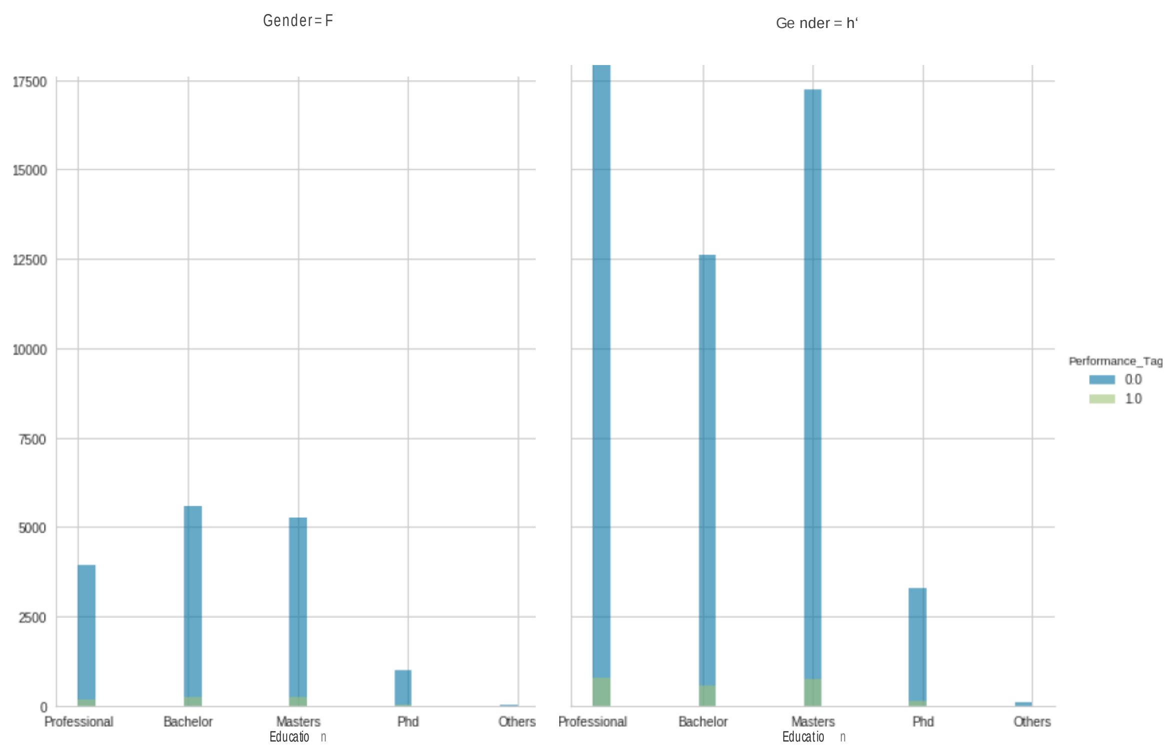


Per formanc e_T ag	6.0	1.0	perc
E due at ton			
Bachelor	16559	7^2	0 D42888
Masters	224B3	99B	B Bd25B2
Others	111	B	0 D67227
Phd	4280	8^	0 D4 1219
Professi on al	23373	1011	0 D41462

- Others has the highest default rate but that could be due to less values in this bin. Among remaining categories, Masters and Bachelor seems to be Defaulting marginally more than the rest of the category.

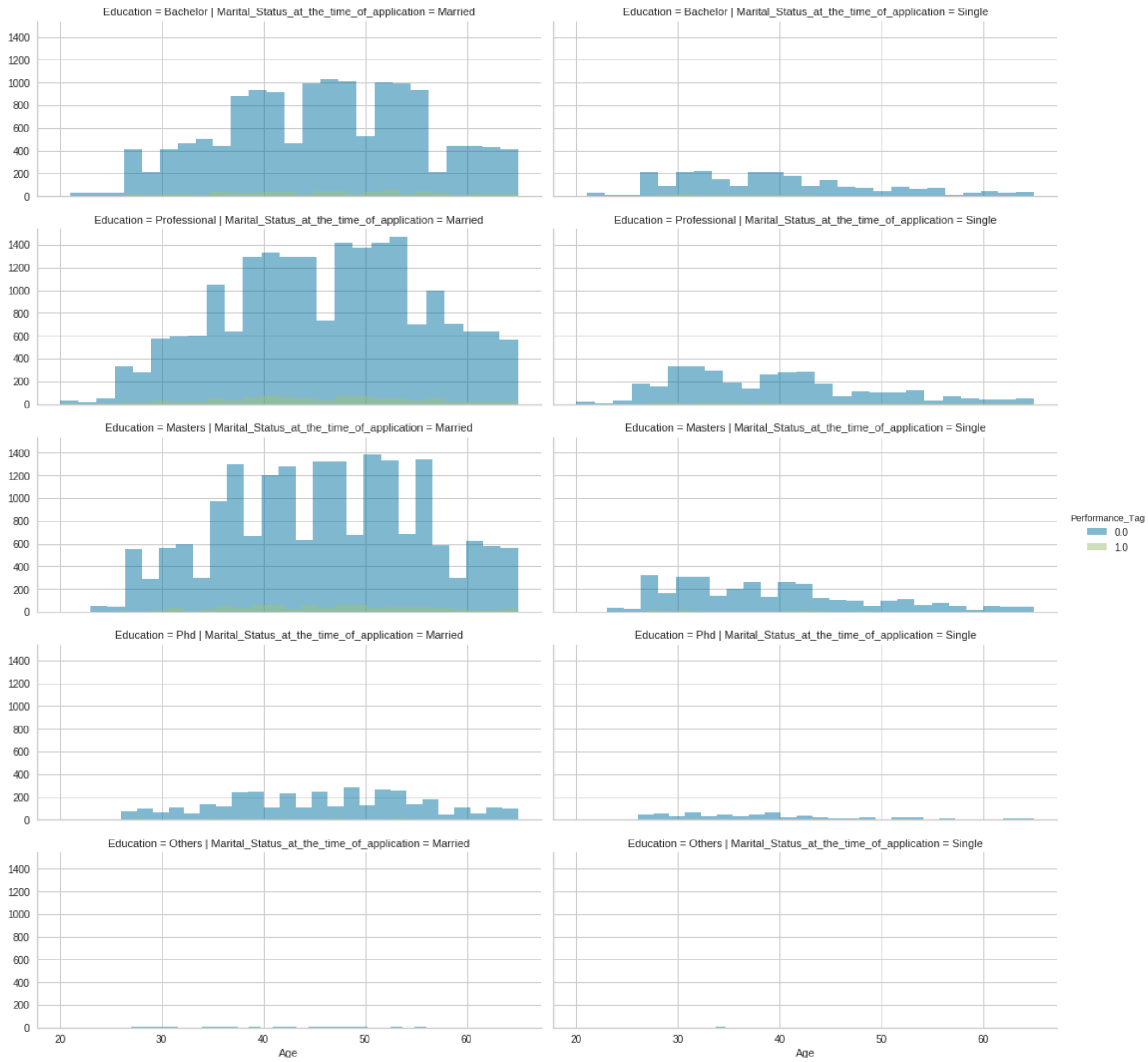


• Masters and Professional education category seems to be defaulting more for Married Applicants.



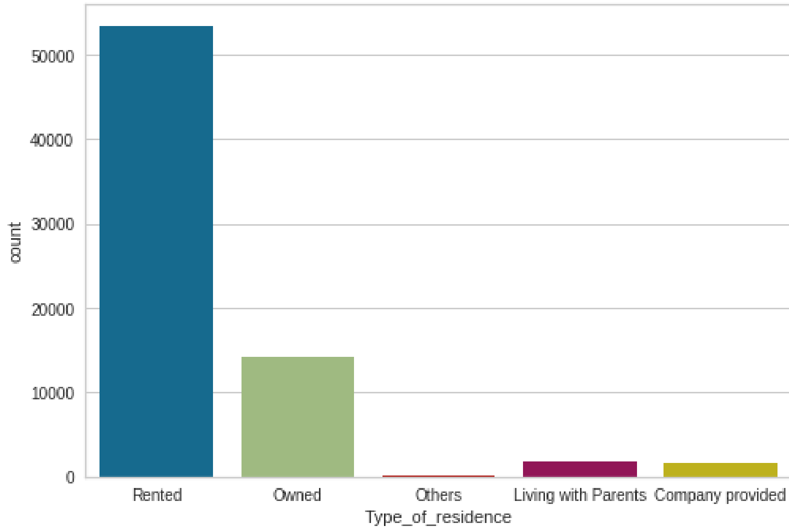
• Professional, Masters and Bachelor degree holders are defaulting more among Male category.

In above plot people in the starting income range in Bachelor, Professional and Masters tend to default more.

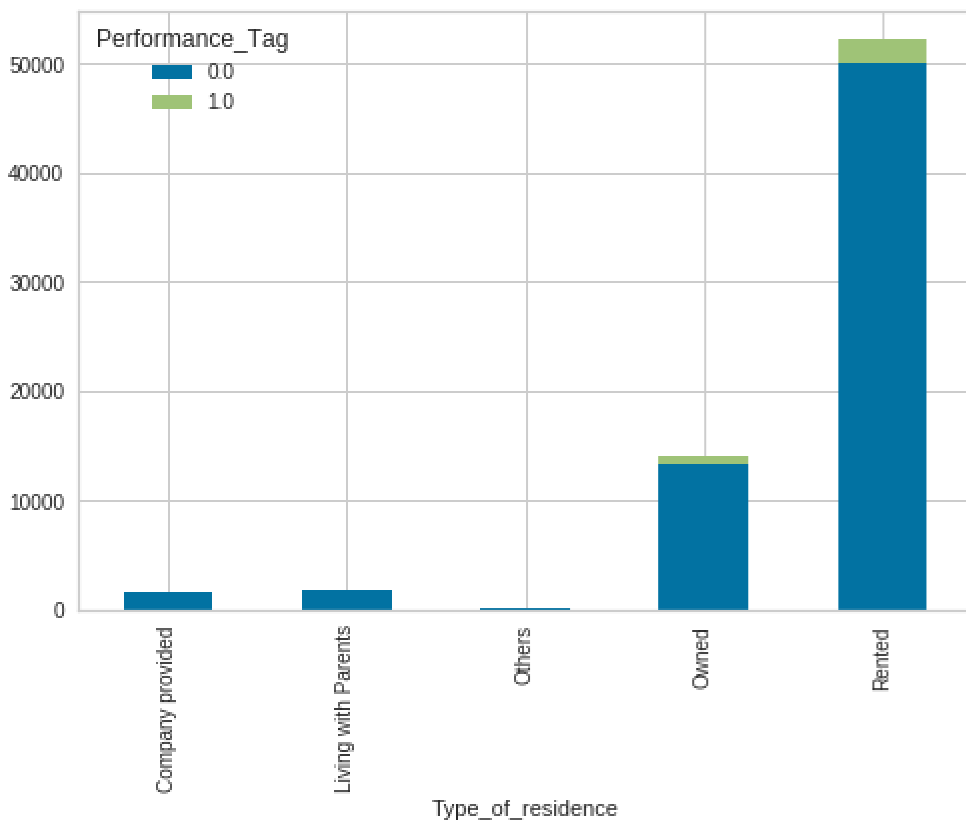


Here we can see that Mid-Aged(between 35-55 age group) professional, Bachelor and Masters who are also married seem to be defaulting more.

➢ Type Of Residence:

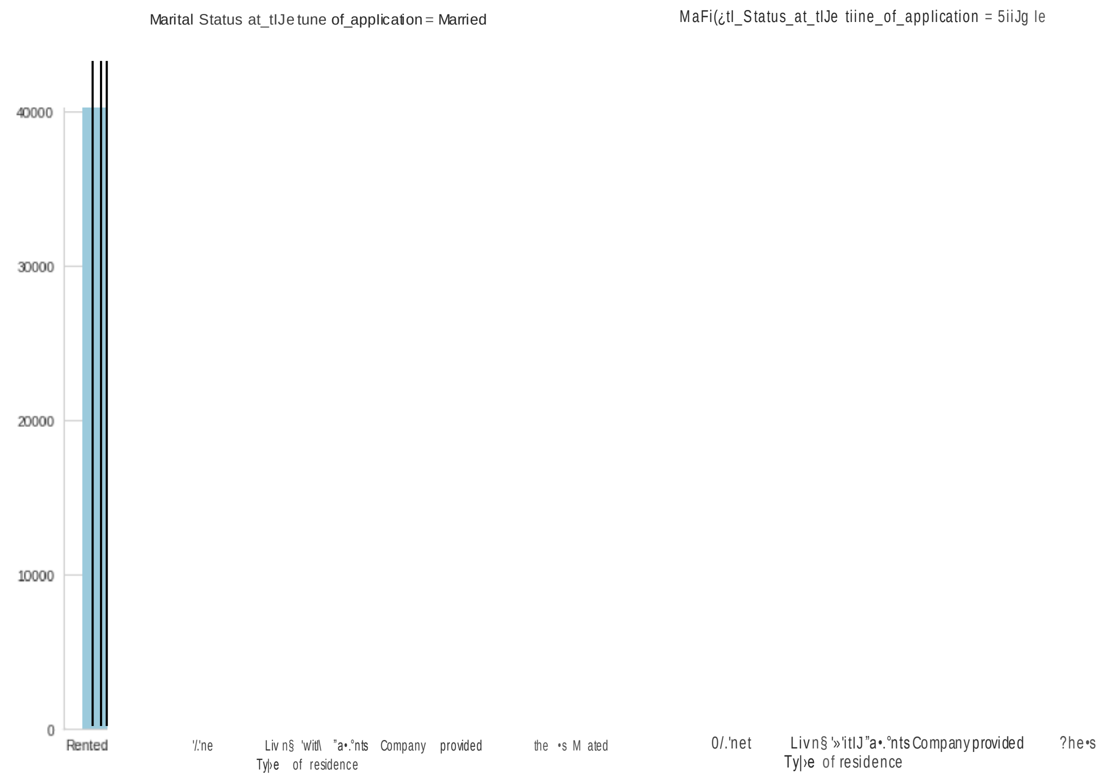


- People who have rented accomodation makes majority of the applicants followed by Owned accomodation.

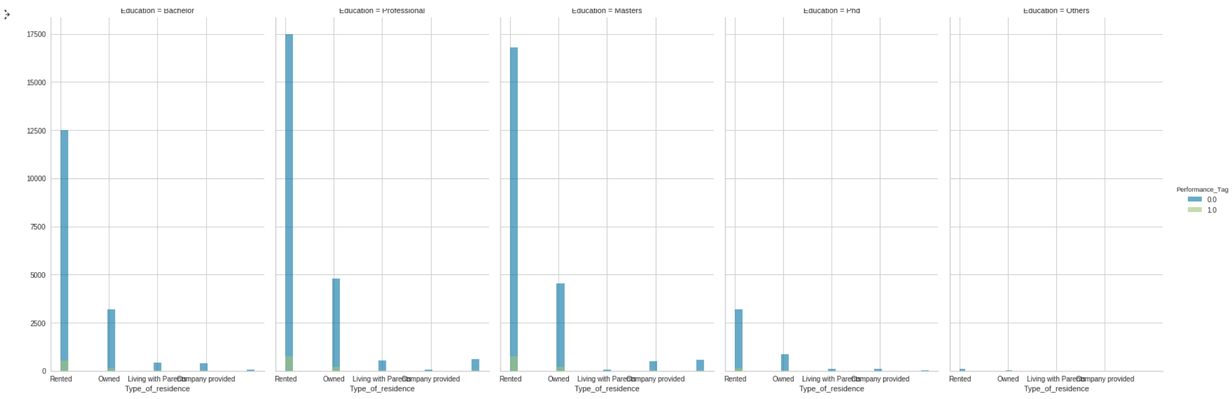


Performance_Tag	0.0	1.0	perc
Type_of_residence			
Company provided	1530	73	0 045540
Liv ing with Parents	1697	80	0.045020
Others	193	5	0025253
Owned	13410	593	0.04234 8
Rented	50081	2197	00M2025

- Company provided accomodatians has the highest default rate but that could be due to less values in this bin. Among remaining categories, Rented and Others seems to be Defaulting marginally more than the rest of the category.

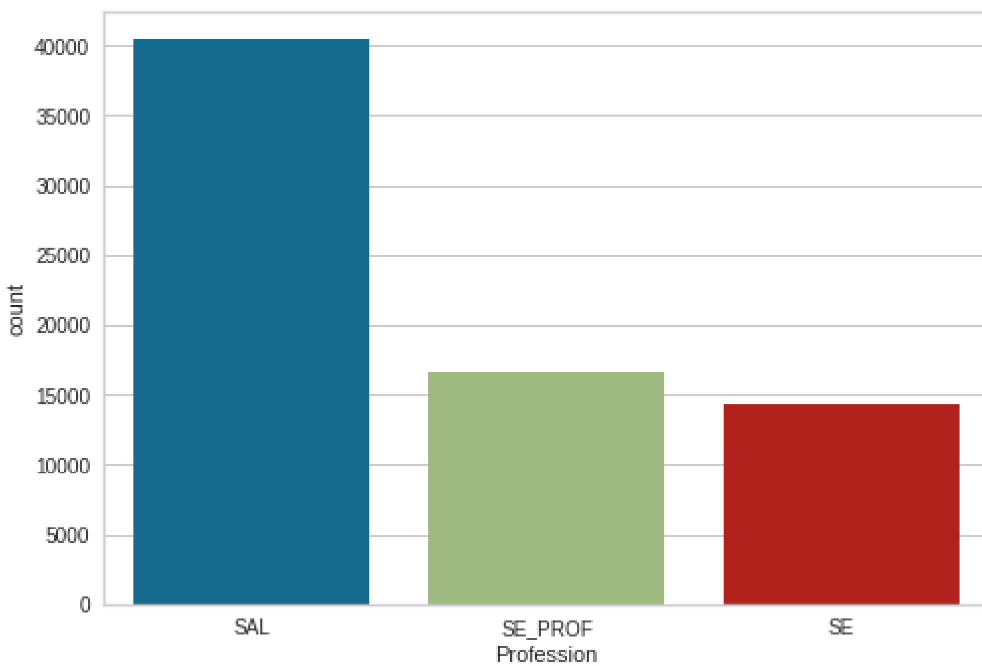


- Rented and Ow ned category seems to be defaulting more far Married Applicants.

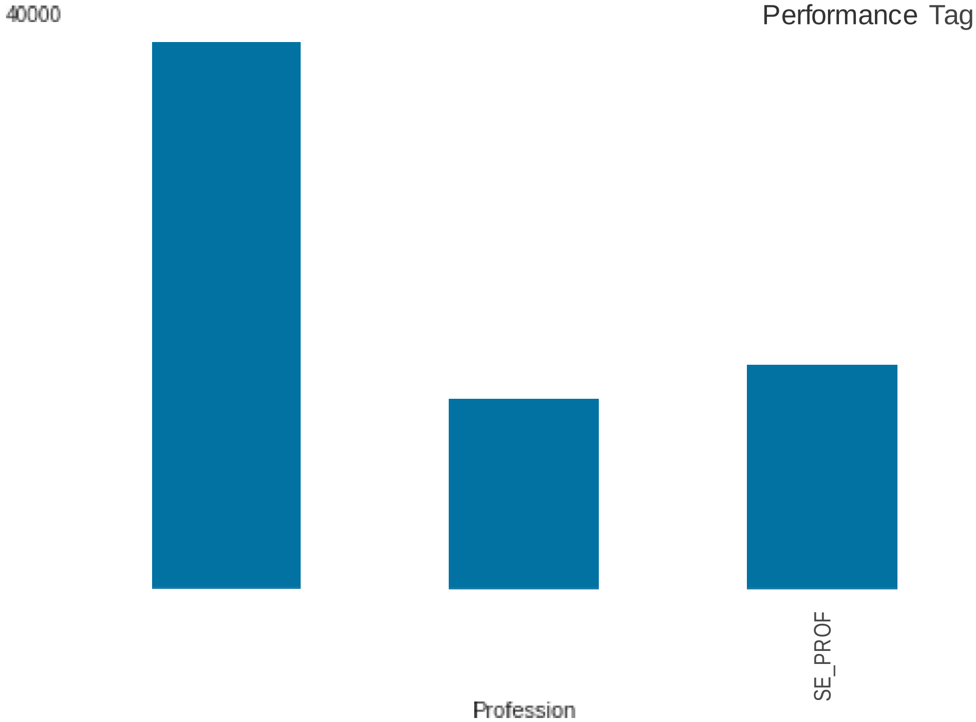


- **Professional**, **Masters** and **Bachelor** degree holders are defaulting more among the Rented category.

➢ **<u>Profession:</u>**

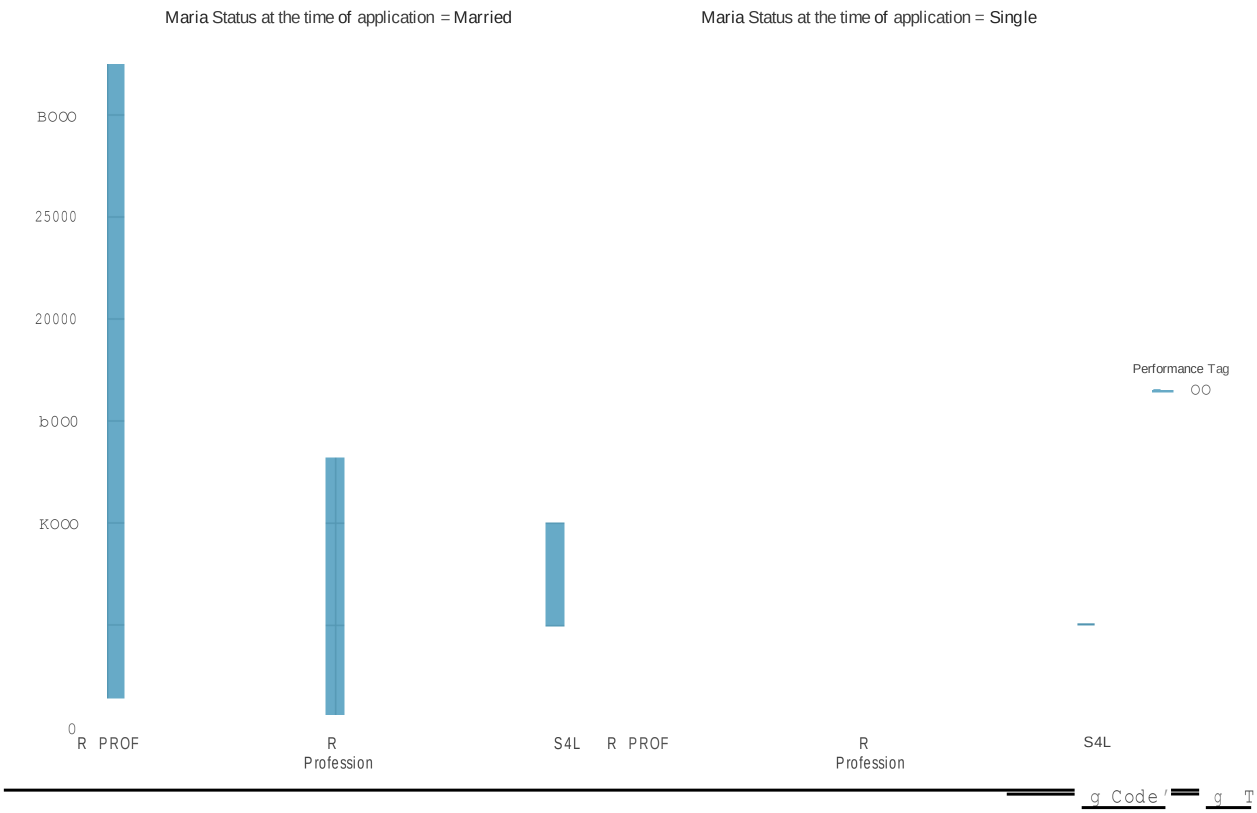


- Mostly Salaried people are applying for the credit card.

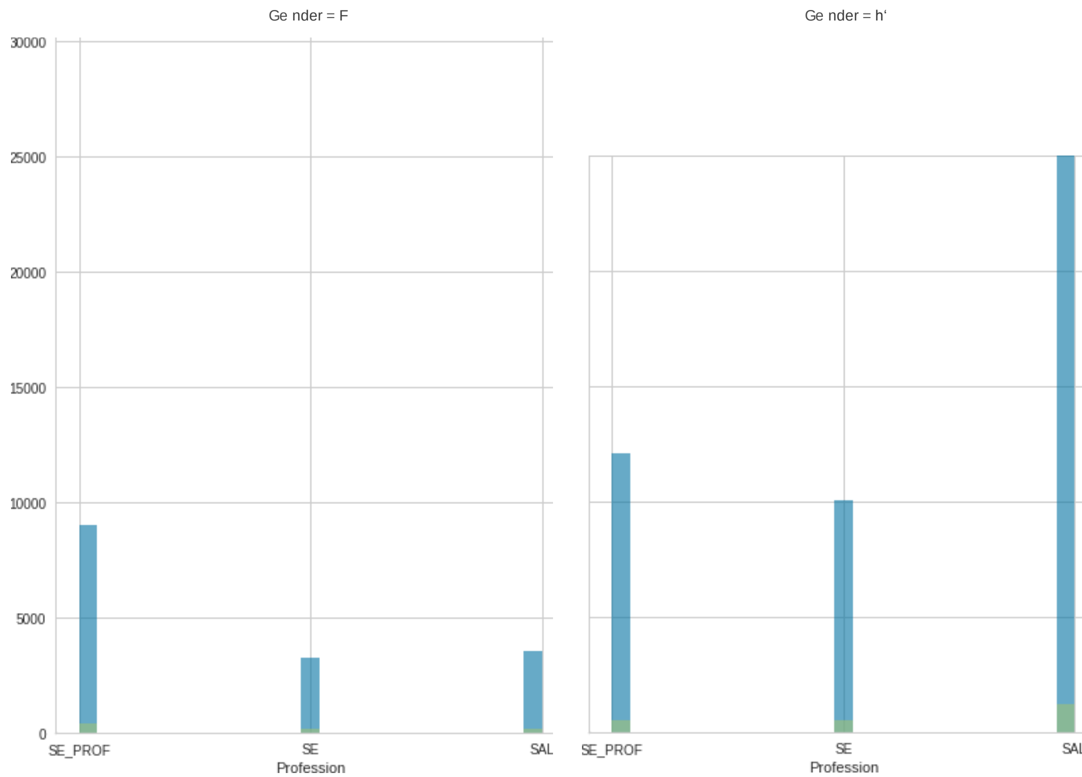


Performance_Tag	0.0	1.0	perc
Prodession			
SAL	38042	1629	0.041063
SE	13285	642	0.046098
SE_PROF	15579	677	0.041646

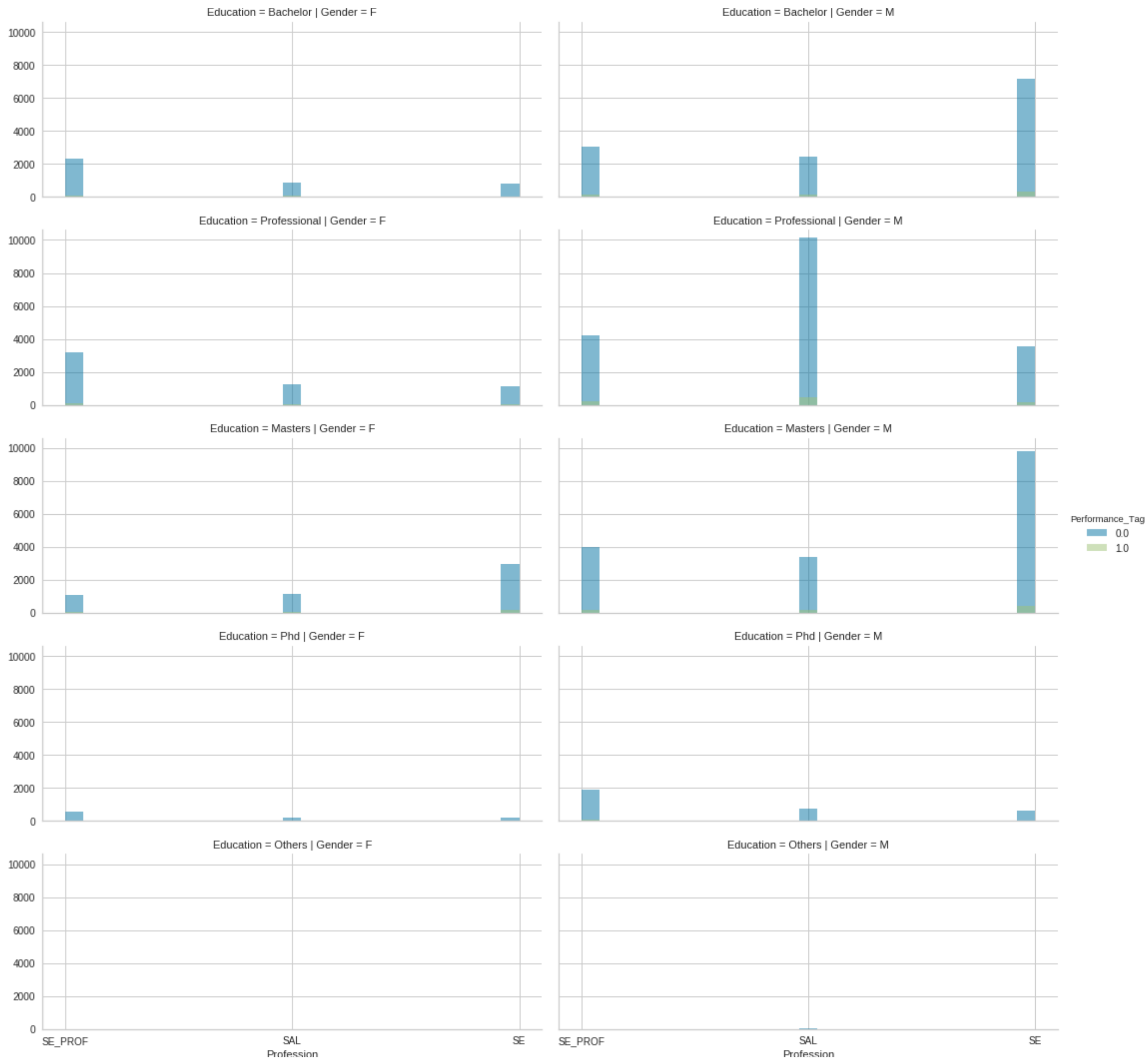
- SE has the highest default rate but that could be due to less values in this bin.



• S1PR0F category seems to be defaulting more for Married Applicants.



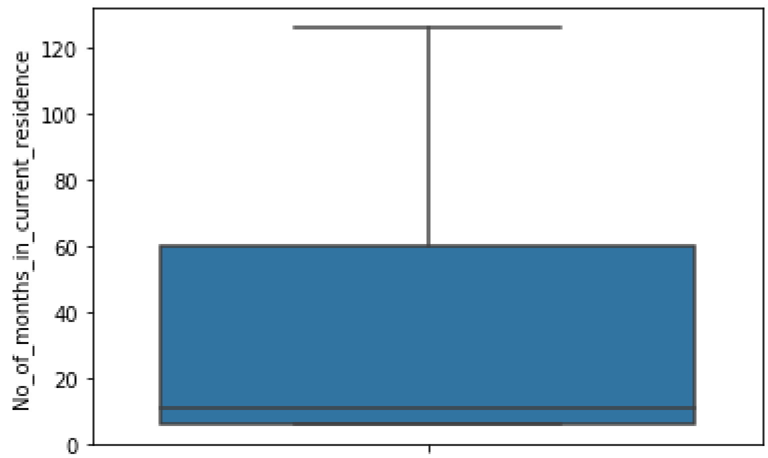
- SAL category are defaulting more among Male category.

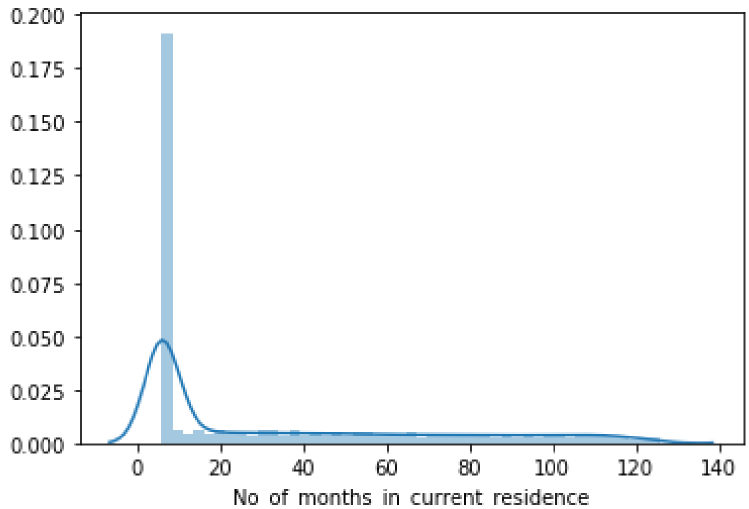


Male, Professional education degree, Salaried are defaulting more than rest of the categories.

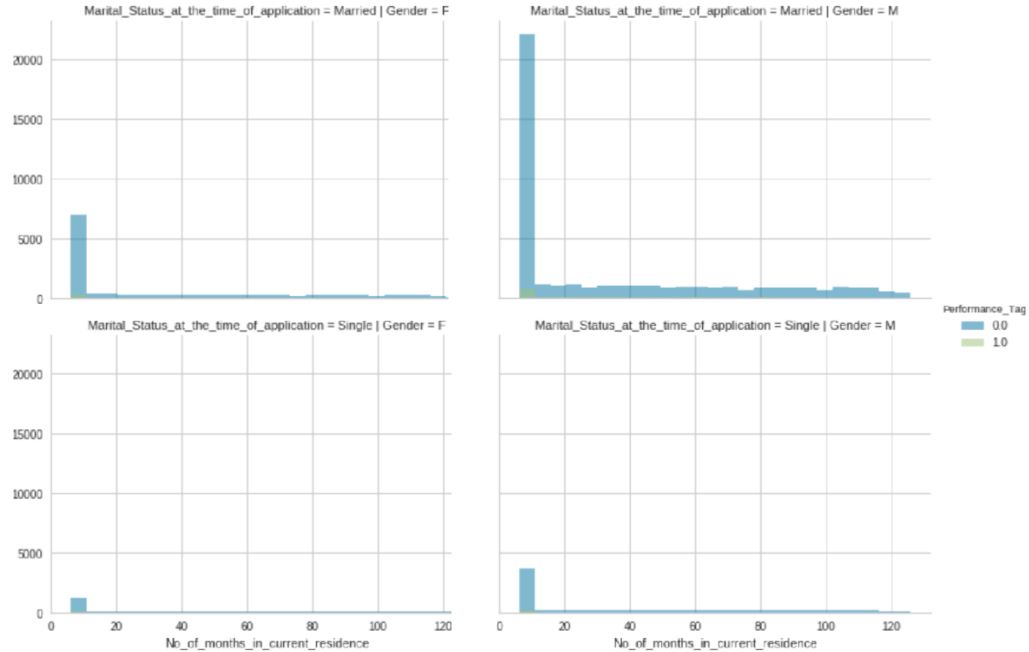
Male, Masters degree holders, SE category also seems to be dafaulting as per the plot. Rest of the Defalters are scattred here and there across different combination of groups.

➢ **No of Months In Current Residence:**

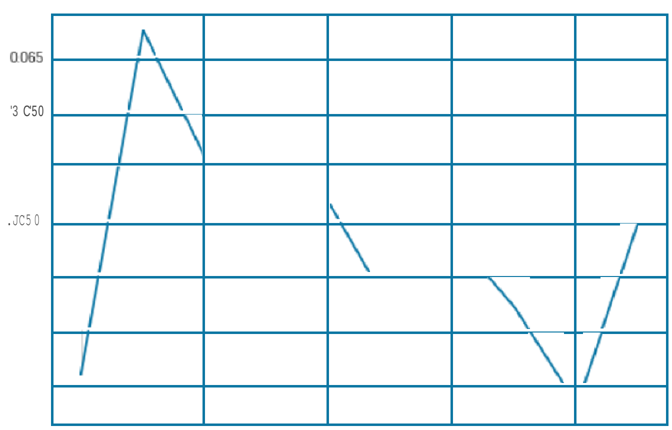




- Looks like No_of_months_in_current_residence is following power law distribution.

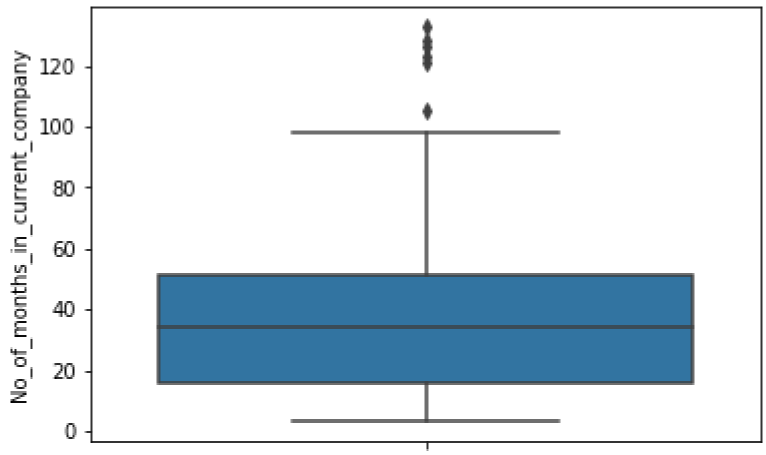


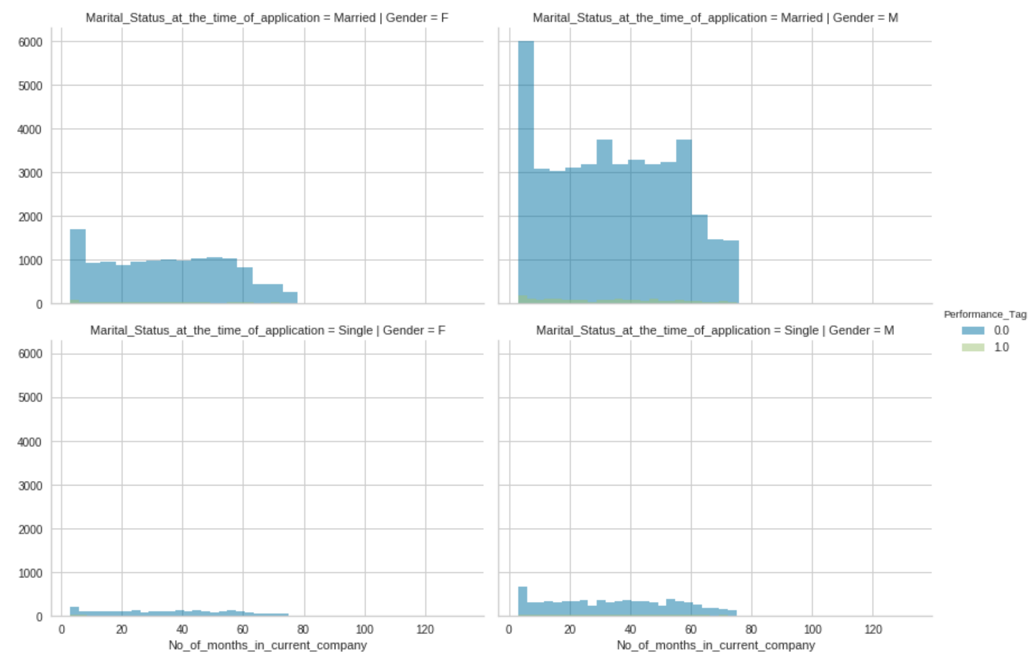
- We can see very clearly see that people who stayed less no. of months in their current accomodation tend to have high default rate.



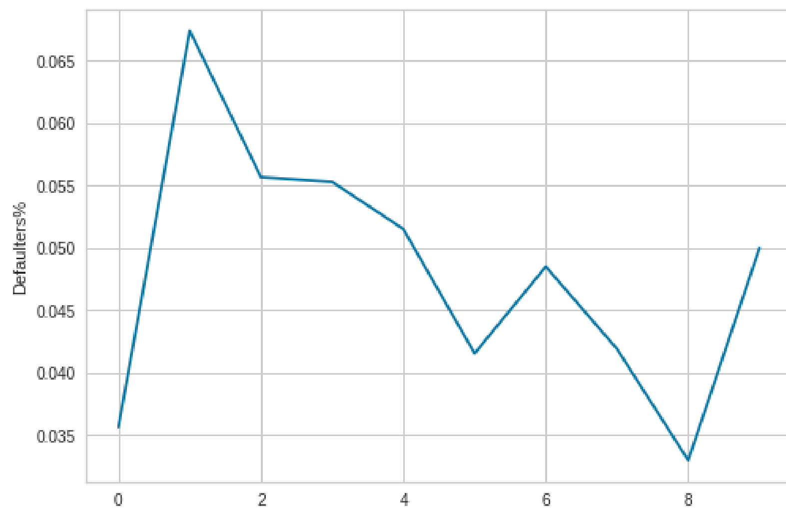
- People in 5-18 months period seem to be defaulting more then people in last category(114126 months)

➢ No Of Months In Current Company:



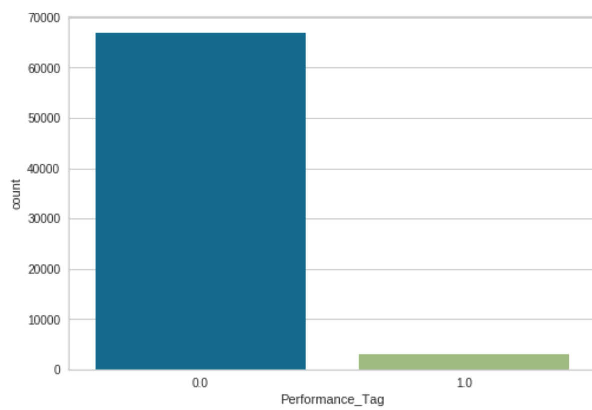


- We can see very clearly see that people who stayed less no. of months in their current company tend to have high default rate. these people are also from Male and Married category.



- People in 6-20 months period seem to be defaulting more then people in last category(62-133 months)

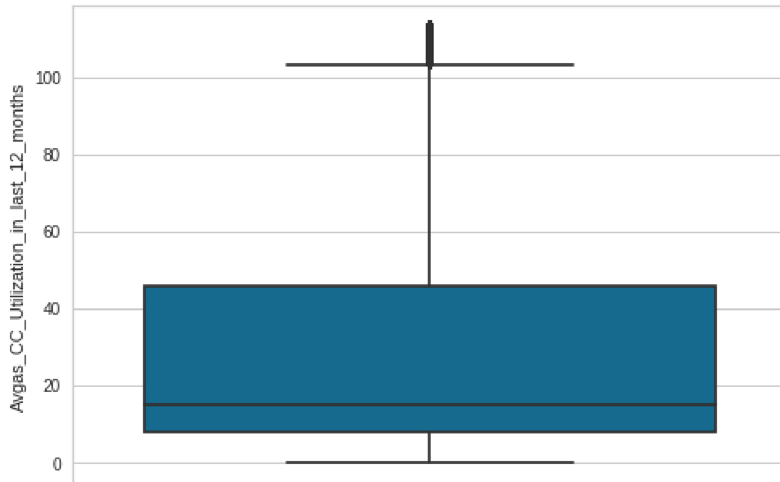
Performance Tag:

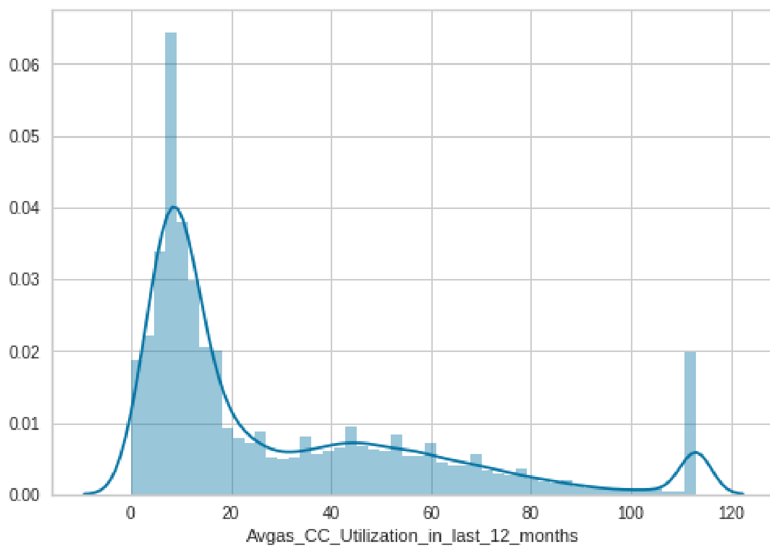


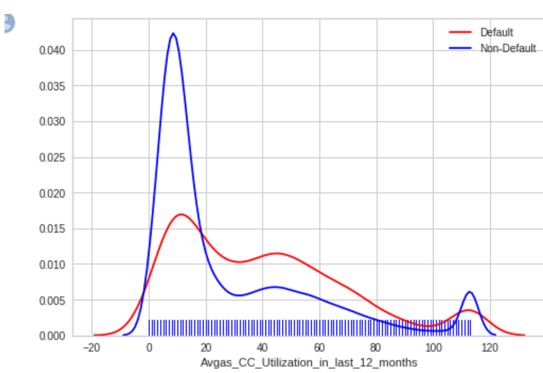
- There's high data imbalance between the two categories.
- Removing Null values from the Performance Tag as there's no way to validate this Data. In ideal situation we could have circled back to the business to double check this data.
- We will validate the performance of these Null rows that are removed now on the Final Model to see what the outcomes are. We will consider these applicants as the ones who would default and were unfit for the Card.

4) <u>Exploratory Data Analysis -- **CREDITBUREAU** Dataset:</u>

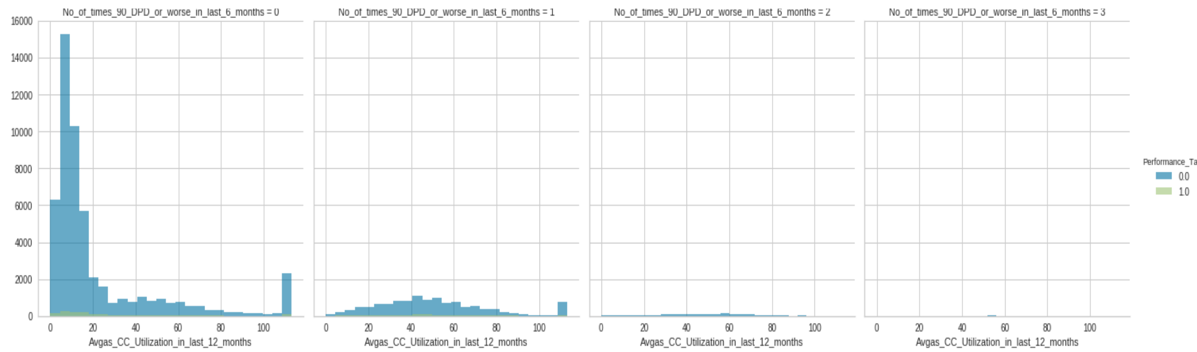
➢ **<u>Avgas_CC_Utilization_in_last_12_months:</u>**



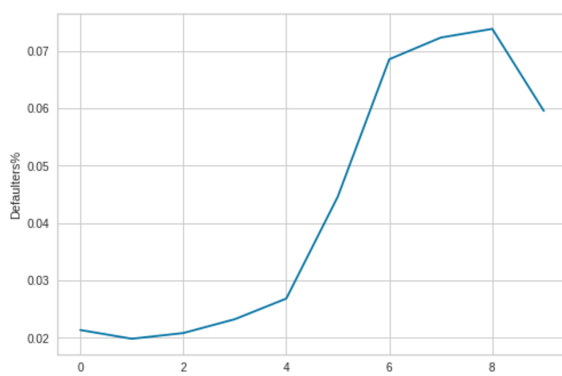




- We can see that users volume is not so high towards high CC utilizations. But between 25 to little over 100 utilizations, Dafualters volume is high.

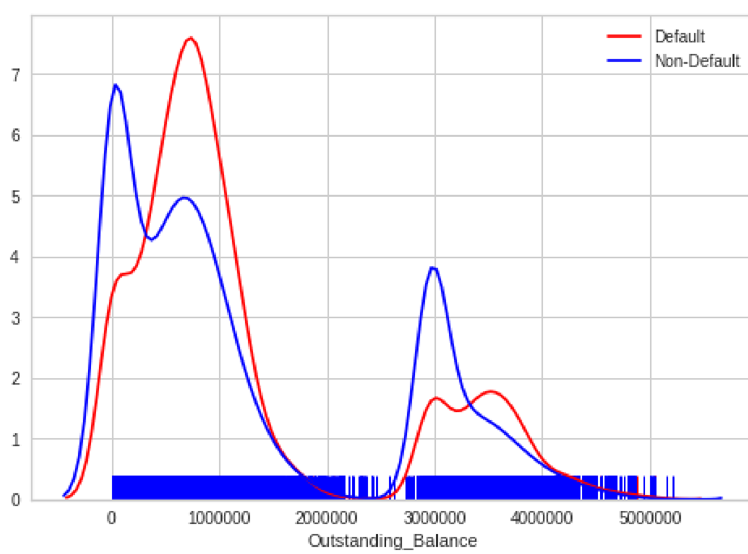


- Based on above plot there's very less data points in the category 2 and 3 of 90 days past due.
- Most people come from 0 and 1 "90 DPD" category

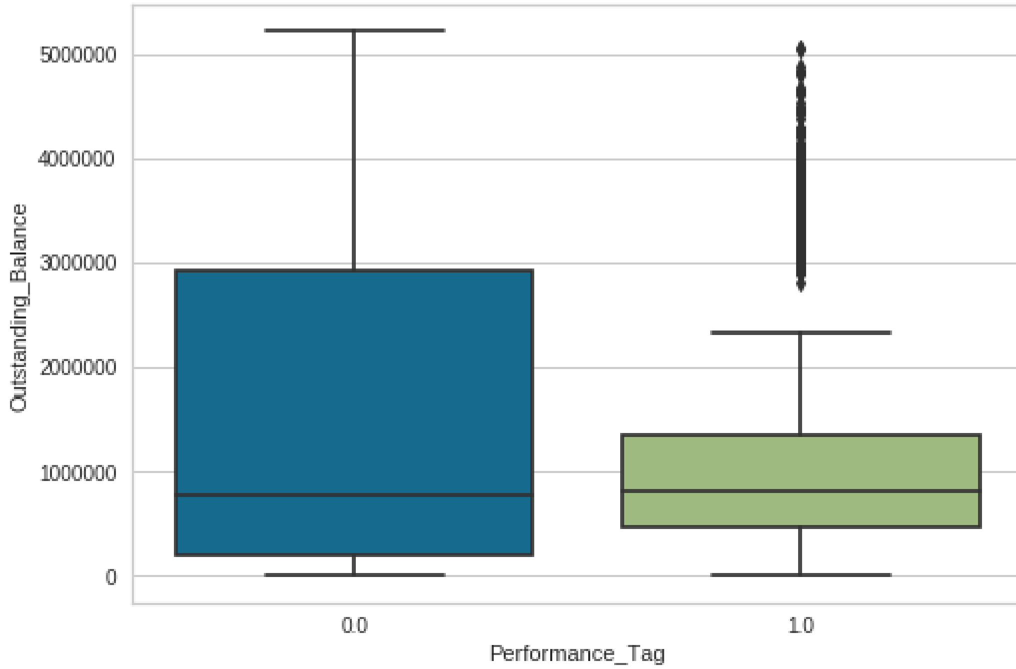


- Looks like the Defaulter % is higher when Credit card utilizations are high

➢ Outstanding Balance:

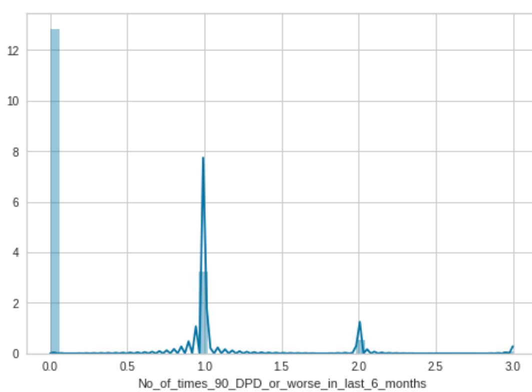


We can see a trend here that people less than 2500000 have more defaulters but after that limit volume of defaulters reduces as compared to the NoN-Default applicants.

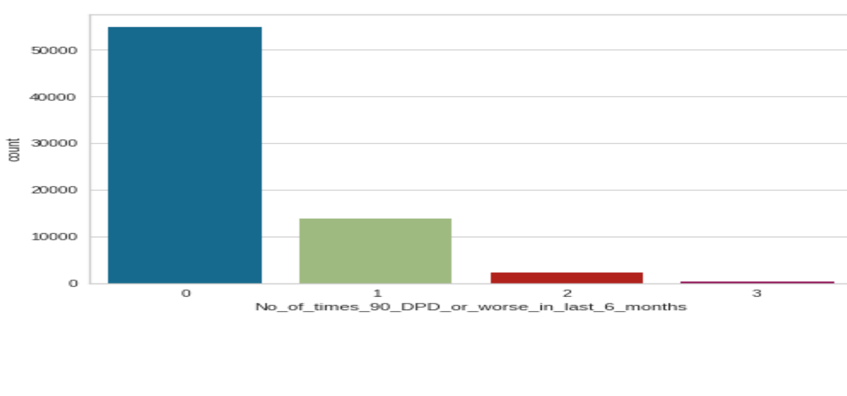


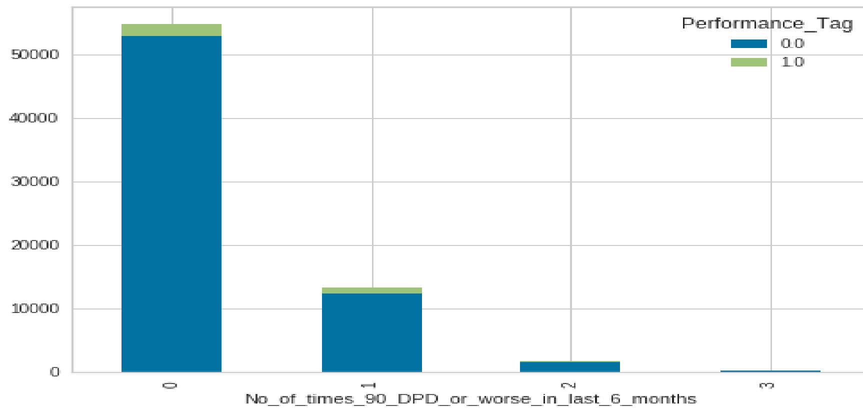
Even here we can see that median is almost same for both the categories.

- **<u>No_of_times_90_DPD_or_worse_in_last_6_months</u>**:



- Max number of people have not gone 90 DPD. Few have gone 1 times and then the count subsequently decreased.

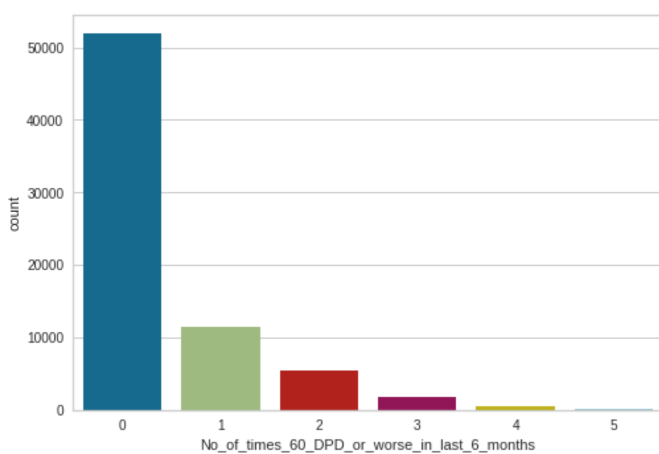


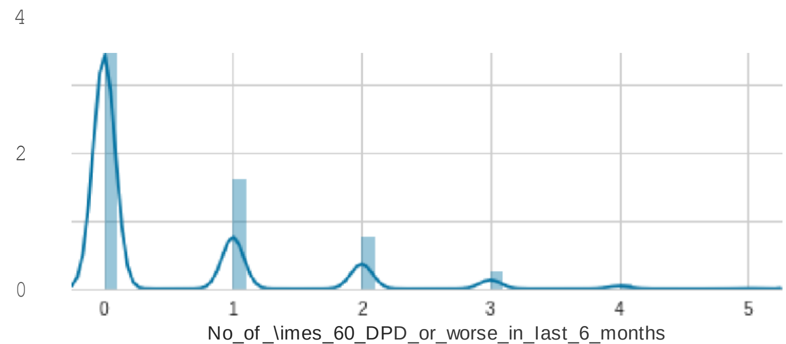


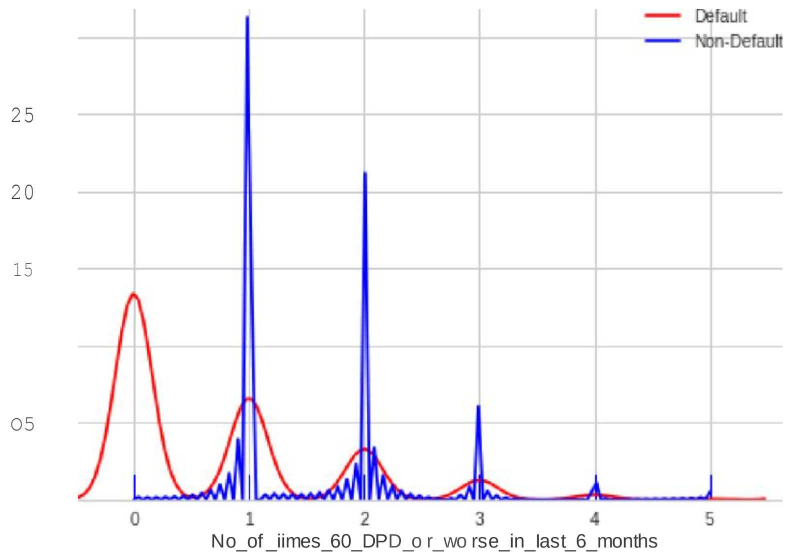
Performance_Tag No_of_times_90_DPD_or_worse_in_last_6_months	0.0	1.0	perc
0	52870	1794	0.032819
1	12248	971	0.073455
2	1616	160	0.090090
3	185	23	0.110577

- As the number of times 90 DPD increase in the last 6 months, the percentage of defaulters have also increased.
- Bank should start taking appropriate actions the moment some crosses 90 DPD for the first time in order to ascertain minimal loss.

➢ No_of_times_60_DPD_or_worse_in_last_6_months:



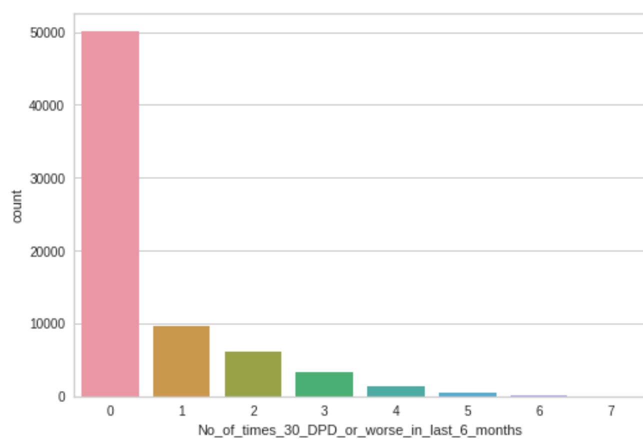

4
2
0
0
1
2
3
4
5
No_of_\imes_60_DPD_or_worse_in_last_6_months

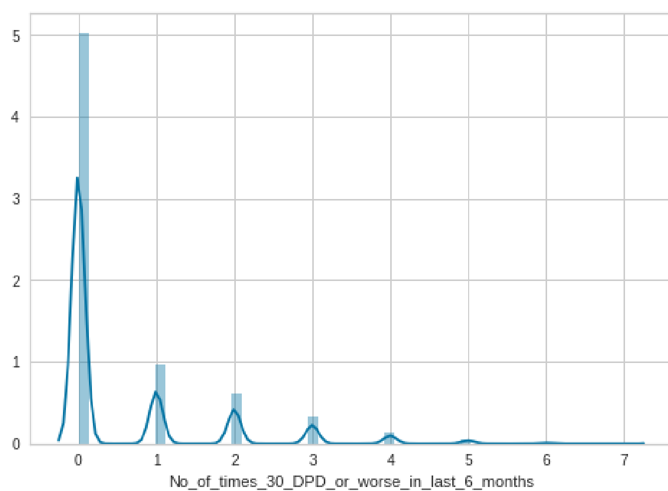

Default
Non-Default
25
20
15
O5
0
1
2
3
4
5
No_of_iimes_60_DPD_o r_wo rse_in_last_6_months

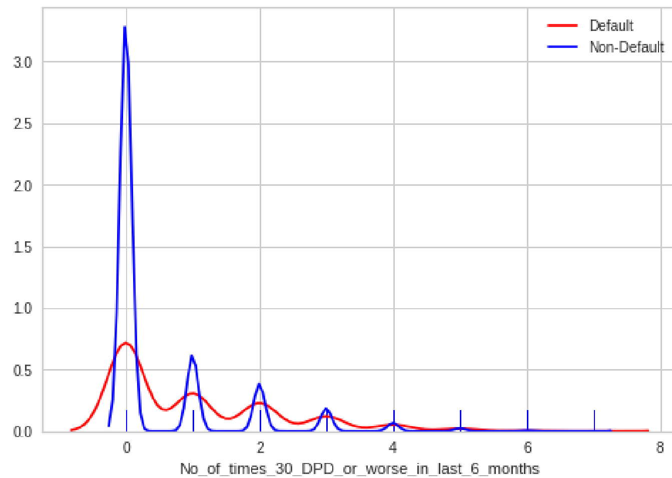
	bins	Defaulters	Count	Non-Defaulters
0	0	1582.0	51869	50287.0
1	1	784.0	11132	10348.0
2	2	389.0	4916	4527.0
3	3	148.0	1469	1321.0
4	4	39.0	411	372.0
5	5	6.0	70	64.0

Based on above plots we can see that as the delinquency count decreases slowly and with that the number of defaulters.

➢ **No_of_times_30_DPD_or_worse_in_last_6_months:**

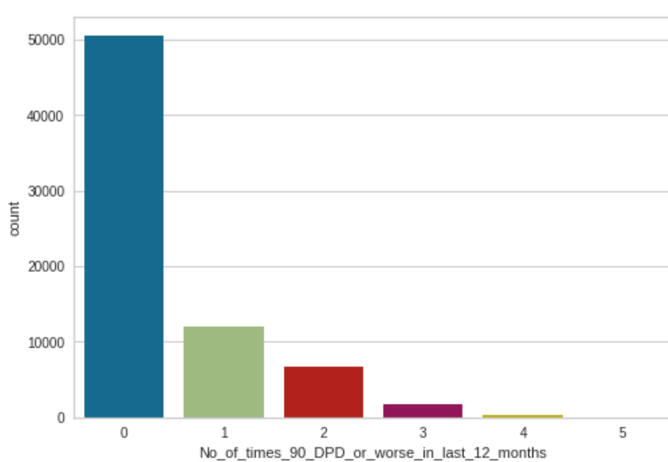


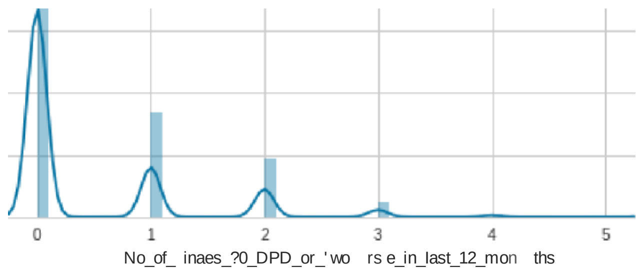


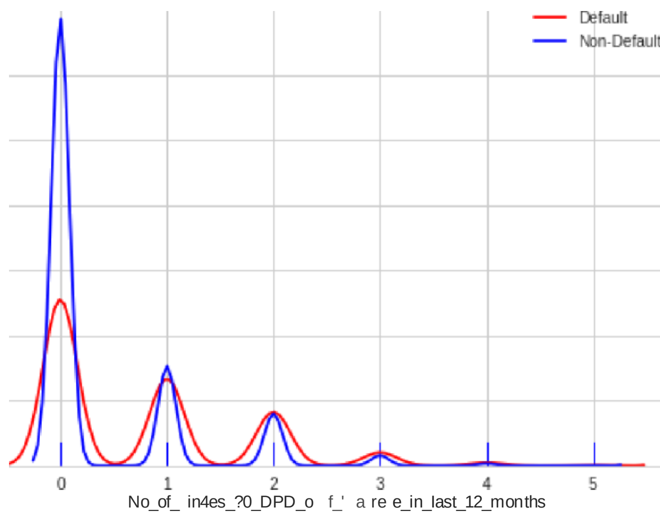


	bins	Defaulters	Count	Non-Defaulters
0	0	1455.0	50097	48642.0
1	1	623.0	9501	8878.0
2	2	466.0	5898	5432.0
3	3	245.0	2829	2584.0
4	4	107.0	1045	938.0
5	5	43.0	386	343.0
6	6	8.0	96	88.0
7	7	1.0	15	14.0

- <u>No_of_times_90_DPD_or_worse_in_last_12_months:</u>

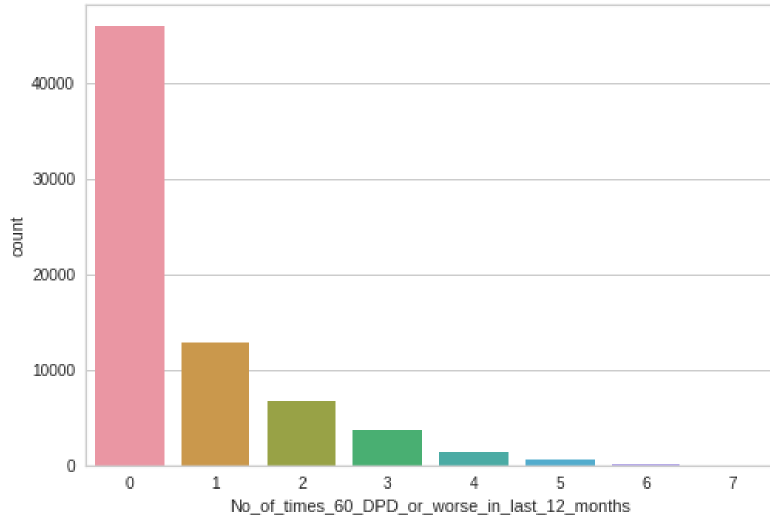


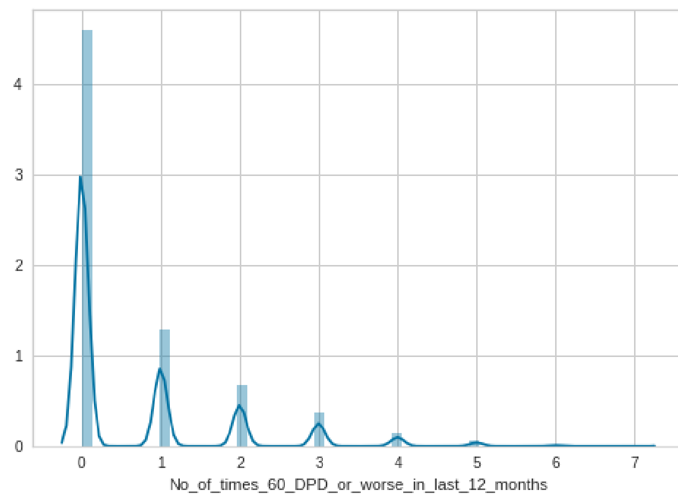


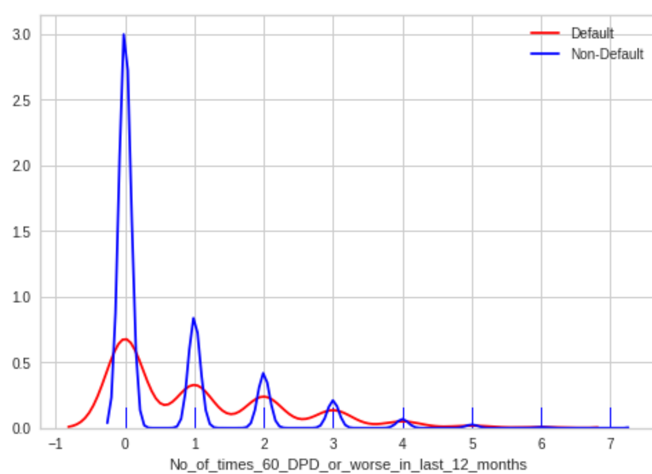


	bins	Defaulters	Count	Non-Defaulters
0	0	1510.0	50492	48982.0
1	1	796.0	11663	10867.0
2	2	489.0	6160	5671.0
3	3	120.0	1244	1124.0
4	4	280	272	244.0
5	5	50	36	31.0

- **No_of_times_60_DPD_or_worse_in_last_12_months :**

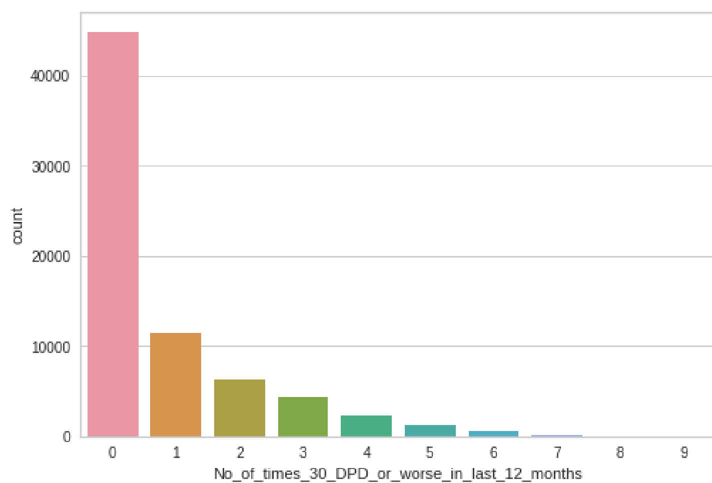


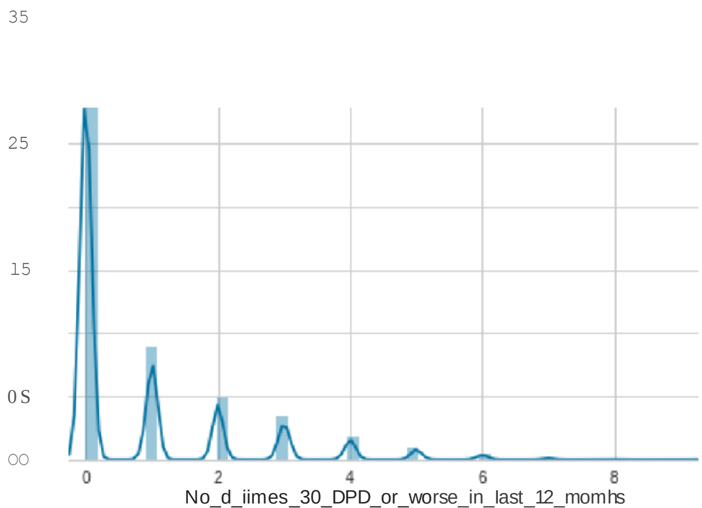


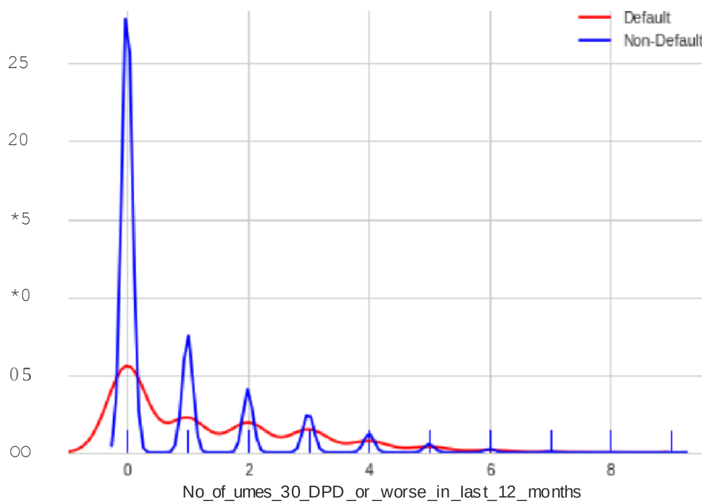


	bins	Defaulters	Count	Non-Defaulters
0	0	1378.0	45867	44489.0
1	1	663.0	12816	12153.0
2	2	483.0	6415	5932.0
3	3	274.0	3205	2931.0
4	4	101.0	1048	947.0
5	5	36.0	398	362.0
6	6	13.0	111	98.0
7	7	0.0	7	7.0

- **<u>No_of_times_30_DPD_or_worse_in_last_12_months:</u>**

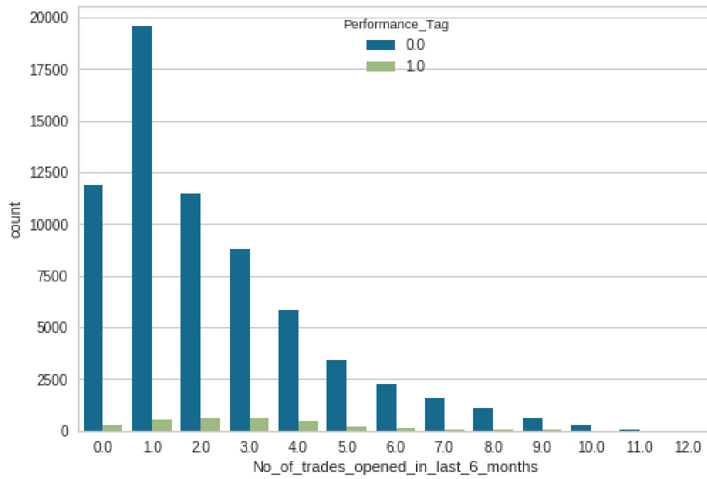


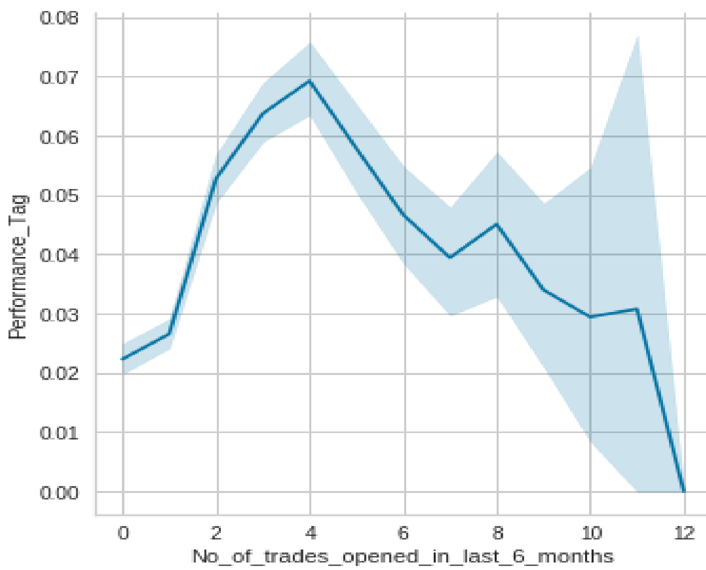


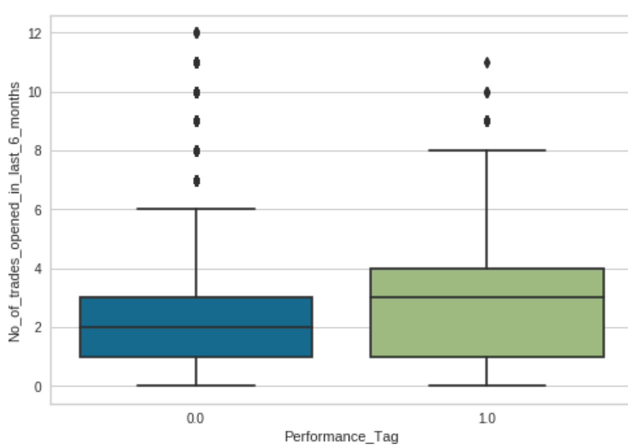


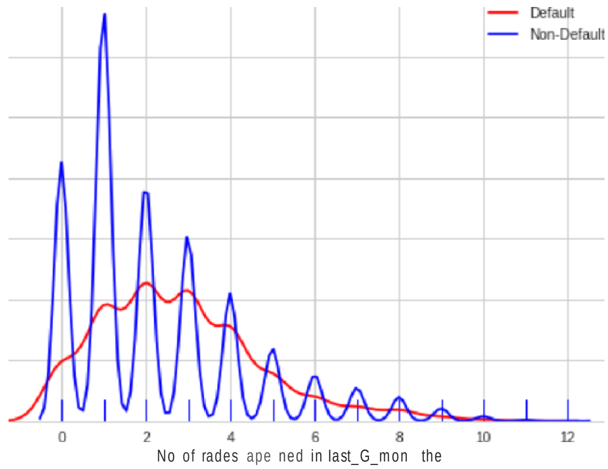
	bins	Defaulters	Count	hon-Defaulters
0	0	1316.0	44B56	435400
1	1	518.0	11474	10956.0
2	2	452.D	6117	5665.0
3	3	349.D	4136	3787.0
4	4	173.0	1924	1751.0
5	5	B9.0	B53	764.0
6	6	3BD	376	33B.0
T	7	11.D	107	96.0
B	B	2.0	23	21 .0
9	9	00	1	1 .0

➤ **No_of_trades_opened_in_last_6_months:**



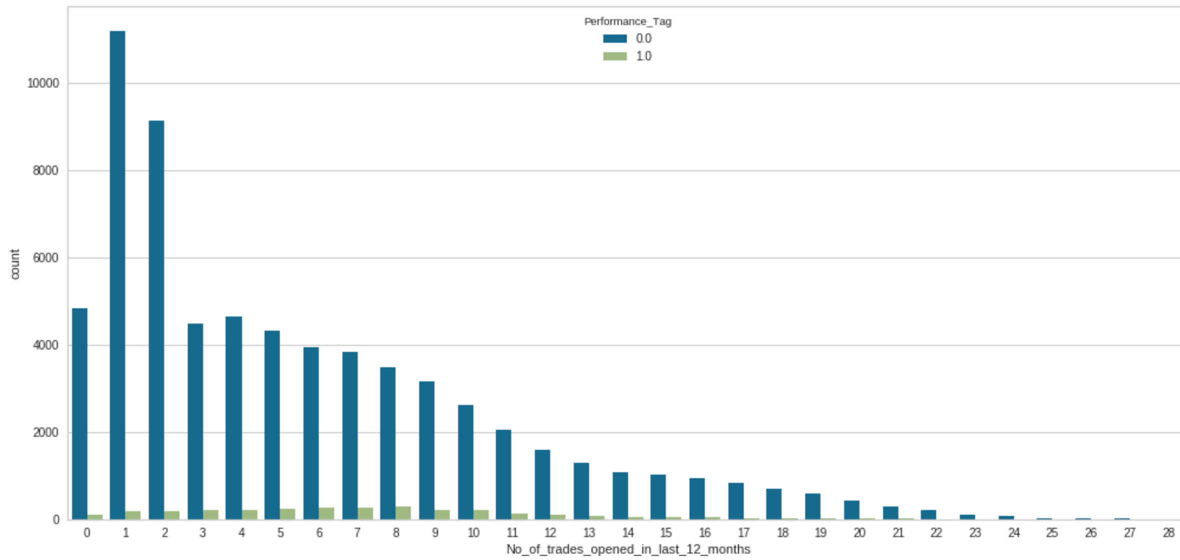


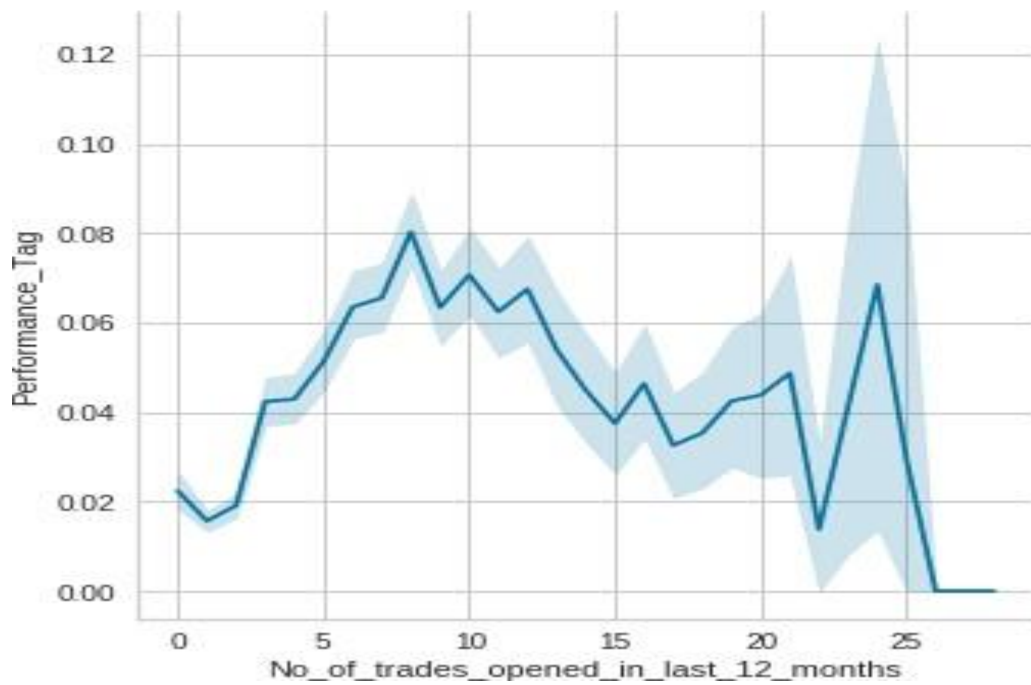


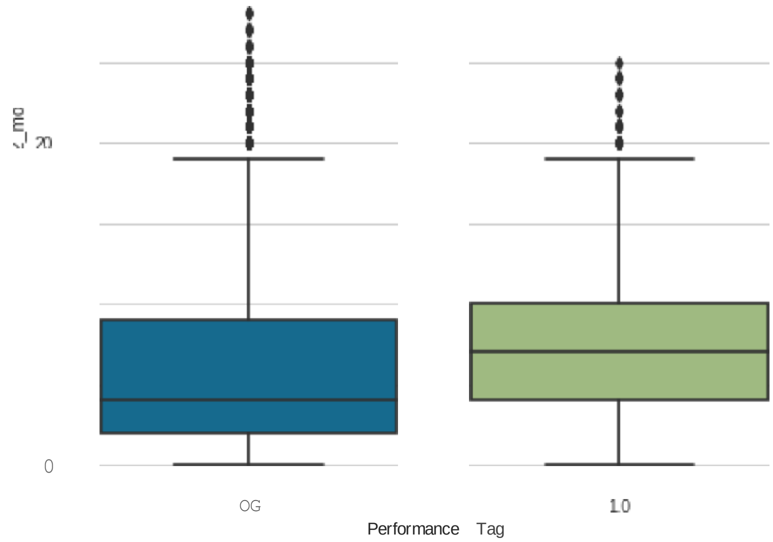


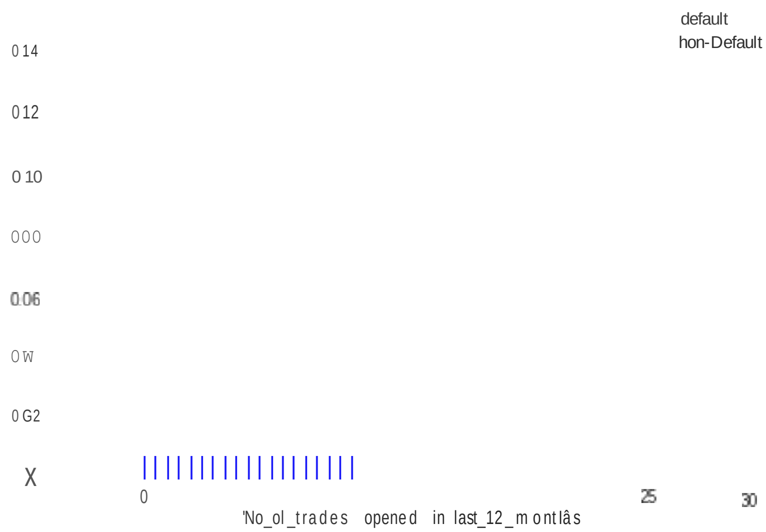
	bins	Defaulters	Count	hon-Defaulters
0	0.0	272.0	1 2193	11921.0
1	1.0	534.0	20121	19587.0
2	2.0	639.0	1 2116	11477.0
3	3.0	599 0	9403	8804.0
4	40	436.0	6297	5861 .0
5	50	212.0	3665	3453.0
6	6.0	109.0	2336	2227.0
T	7.0	65.0	1649	1 584.0
8	80	52.0	1154	1102.0
9	9.0	21.0	618	597.0
10	TOO	70	238	231.0
11	11.0	2.0	65	63.0
12	12.0	00	11	11.0

➢ **No_of_trades_opened_in_last_12_months:**

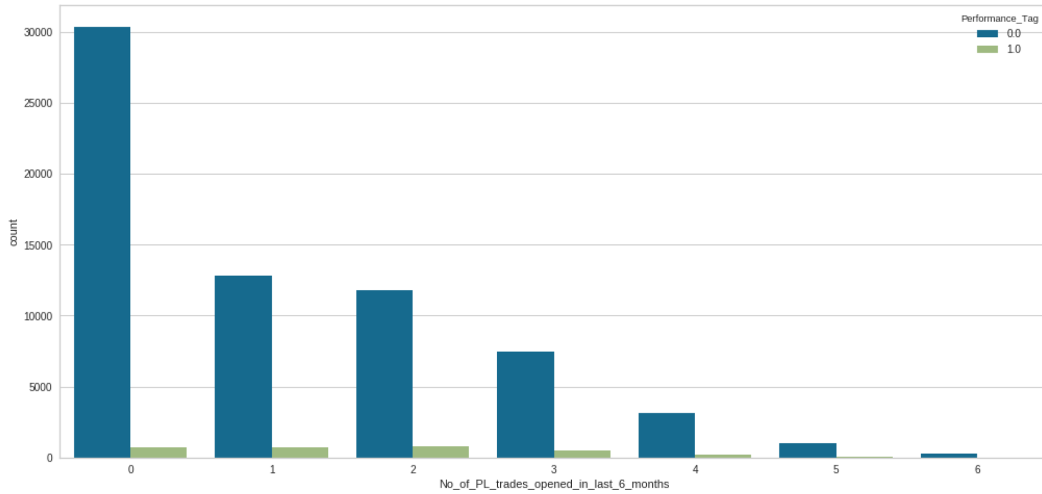


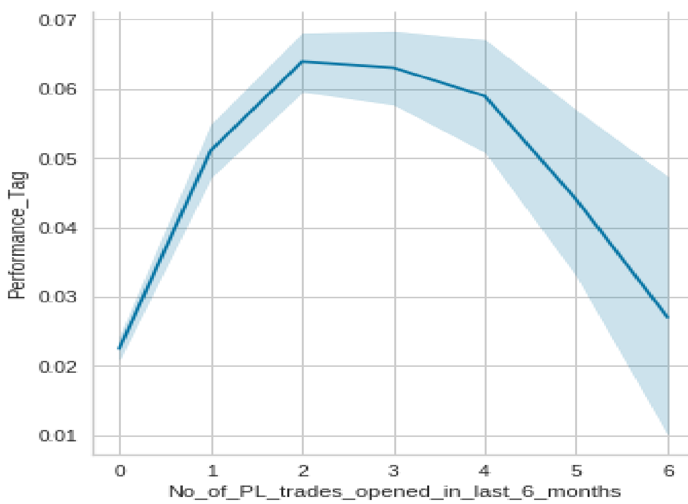


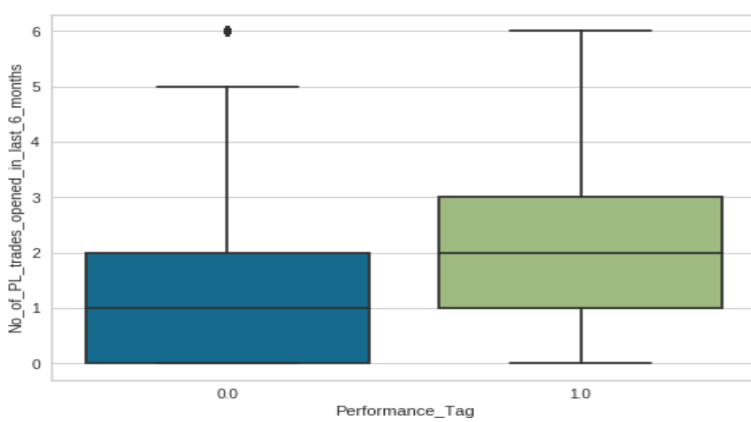

20
0
OG
1.0
Performance Tag

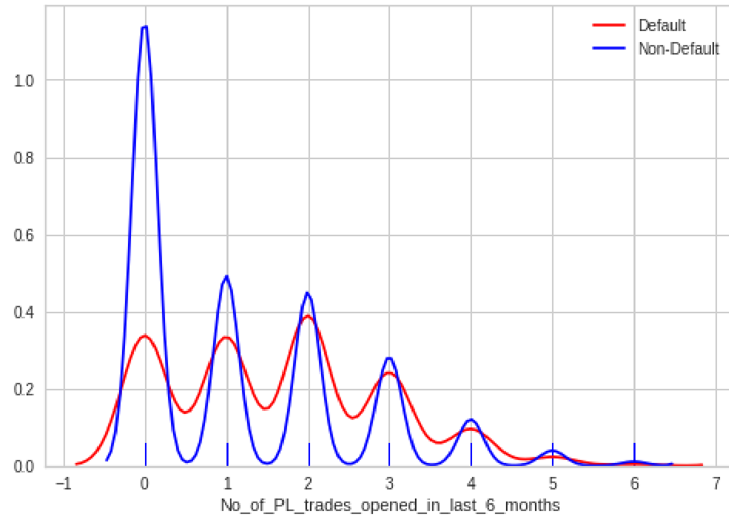

default
hon-Default
0 14
0 12
0 10
0 G2
0
25
30
'No_ol_trades opened in last_12_montlâs

➢ No_of_PL_trades_opened_in_last_6_months:



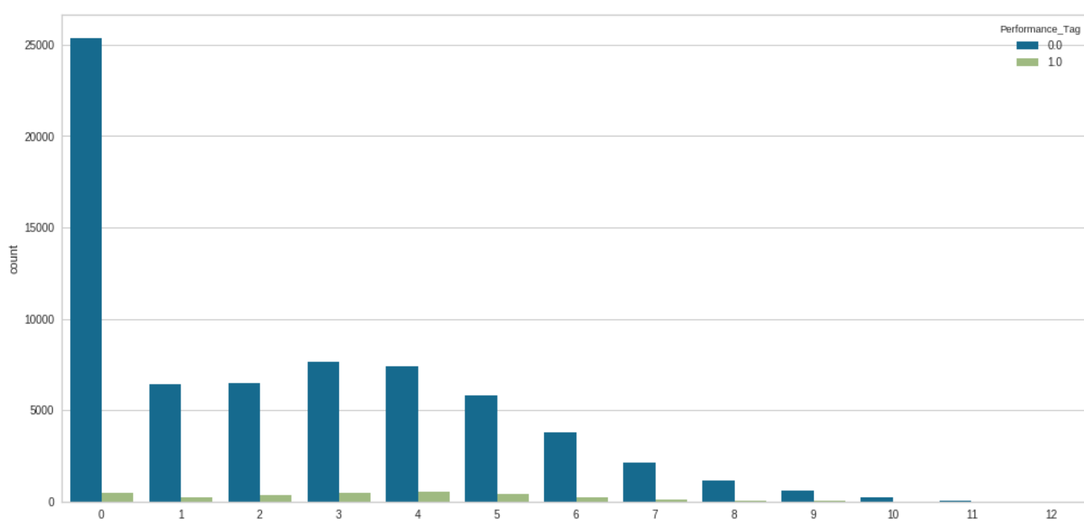


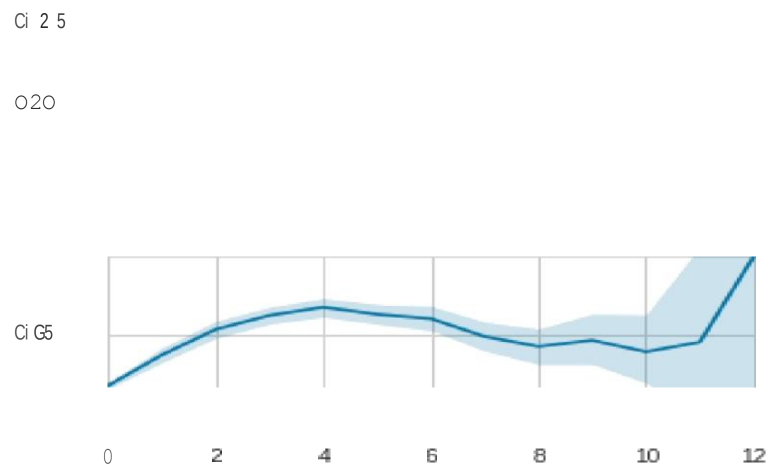


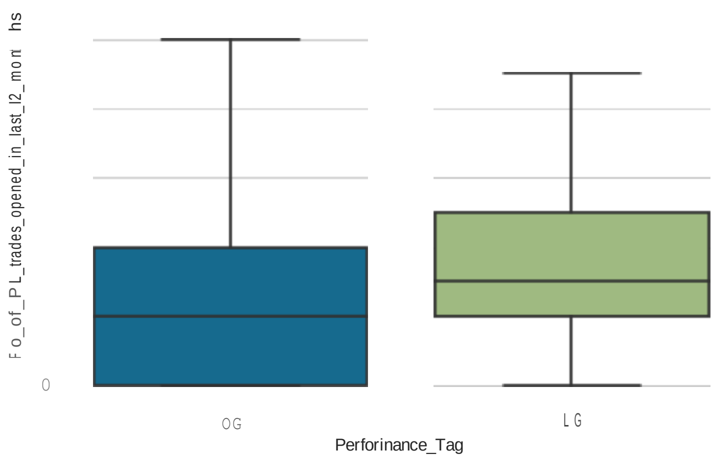


	bins	Defaulters	Count	Non-Defaulters
0	0	699.0	31079	30380.0
1	1	692.0	13547	12855.0
2	2	803.0	12565	11762.0
3	3	501.0	7949	7448.0
4	4	197.0	3341	3144.0
5	5	48.0	1090	1042.0
6	6	8.0	296	288.0

➢ <u>**No_of_PL_trades_opened_in_last_12_months:**</u>

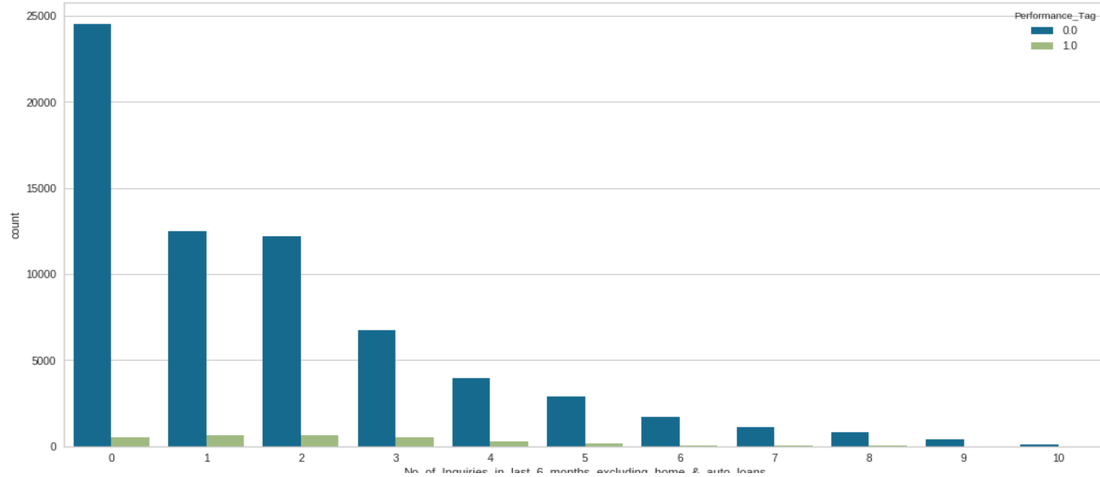


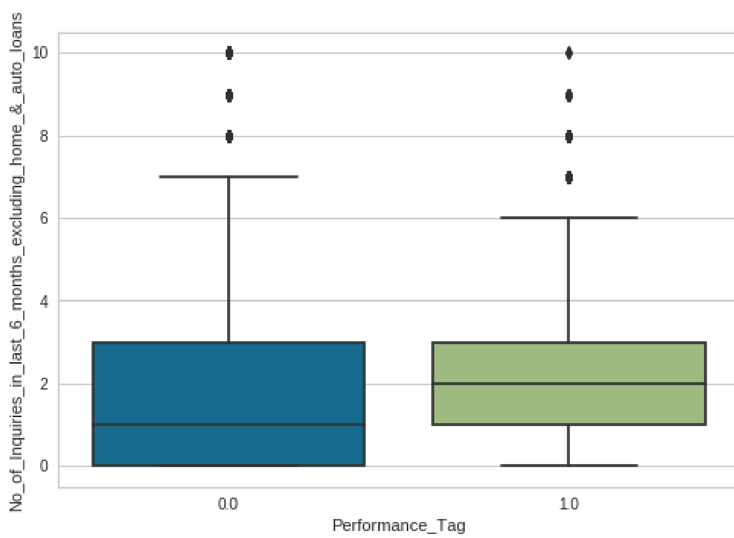


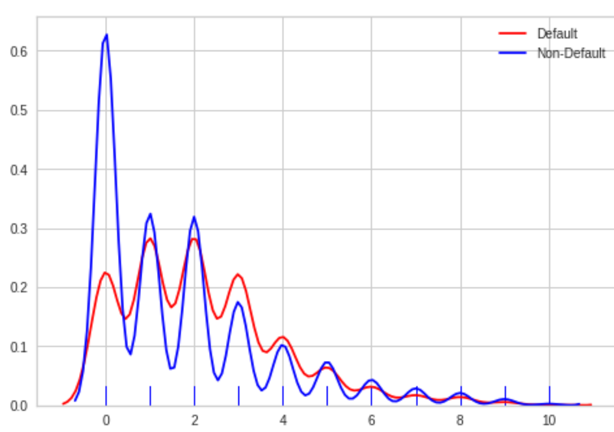


	b1 n s	De-Fa u lt er s	t4on De-Fa u lt er s	
1	1	24 7 . 0	6 64 1	fi D 94 .0
2	2	36 6. 0	6B3O	6d6d.0
3	3	5O8. O	B 13 1	7623 O
4	4	535.0	7903	736B 0
5	5	39 1 . 0	61 B9	579B 0
6	6	2430	40T3	37BO 0
7	7	10 9. 0	2223	2114. O
8	8	5O . 0	11 72	11 22. O
9	9	2 8. 0	60 1	573 0
1 0	1 0	10.0	255	v4&.0
1 1	1 1	30	6 6	63.0
1Z	12	1 . 0	10	90

➢ **No_of_Inquiries_in_last_6_months_excluding_home_&_auto_loans:**

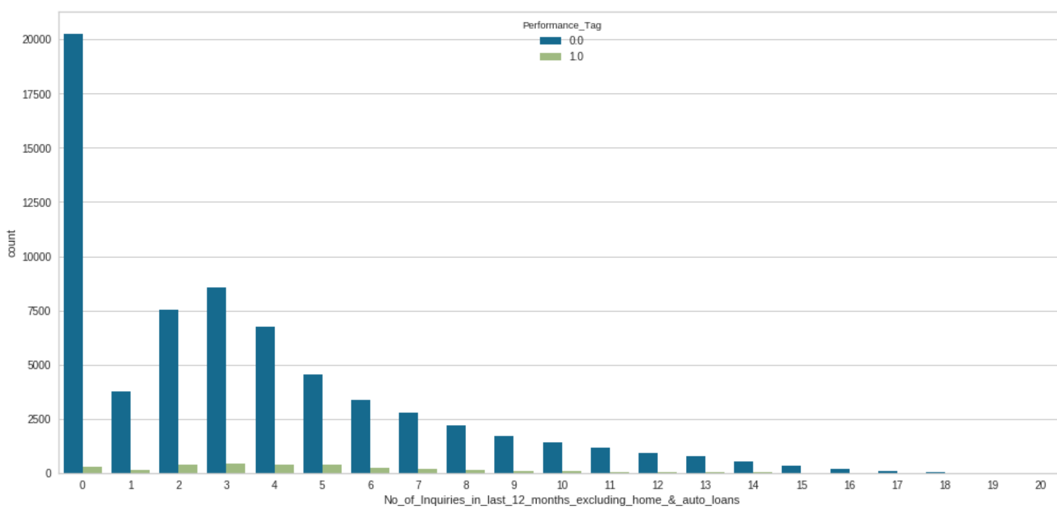


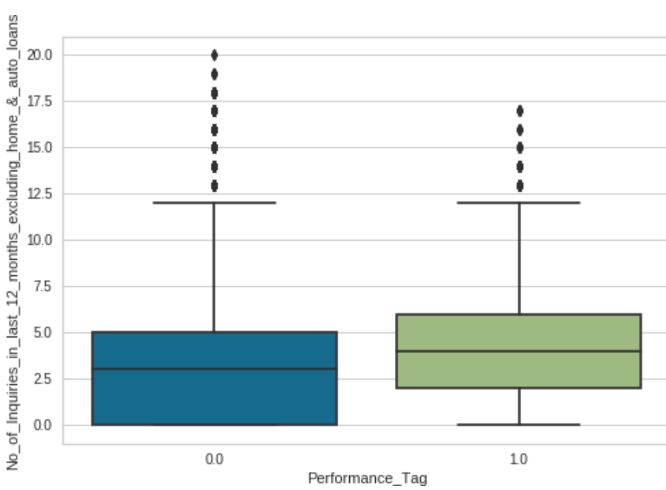


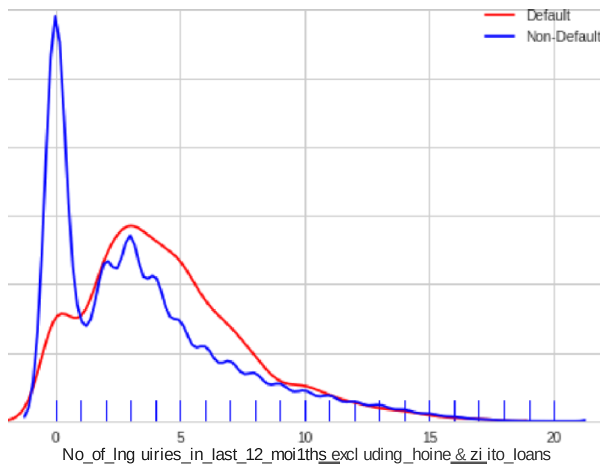


	bins	Defaulters	Count	Non-Defaulters
0	0	527.0	25068	24541.0
1	1	659.0	13176	12517.0
2	2	665.0	12832	12167.0
3	3	517.0	7257	6740.0
4	4	269.0	4248	3979.0
5	5	150.0	3019	2869.0
6	6	73.0	1750	1677.0
7	7	40.0	1149	1109.0
8	8	33.0	835	802.0
9	9	13.0	425	412.0
10	10	2.0	108	106.0

- **No_of_Inquiries_in_last_12_months_excluding_home_&_auto_loans:**

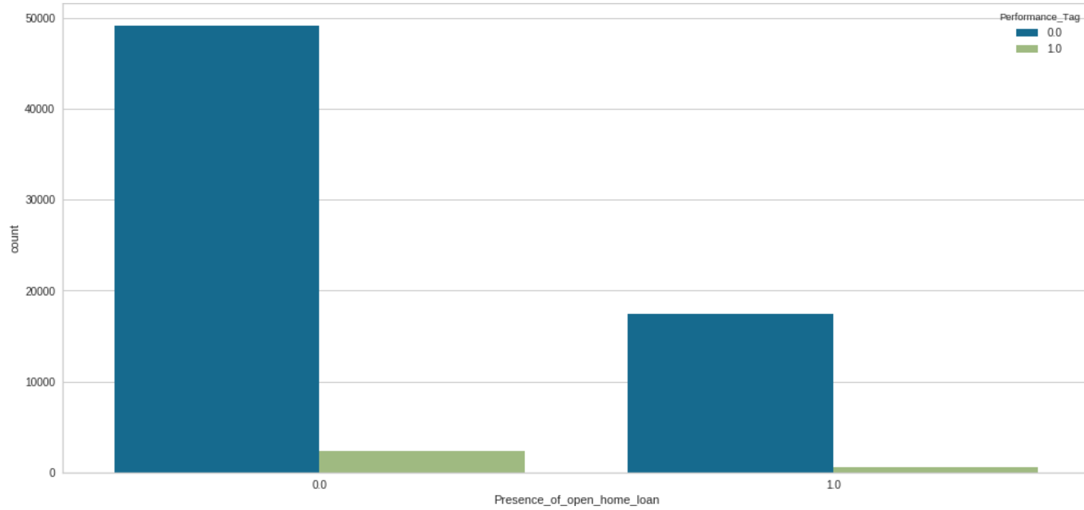






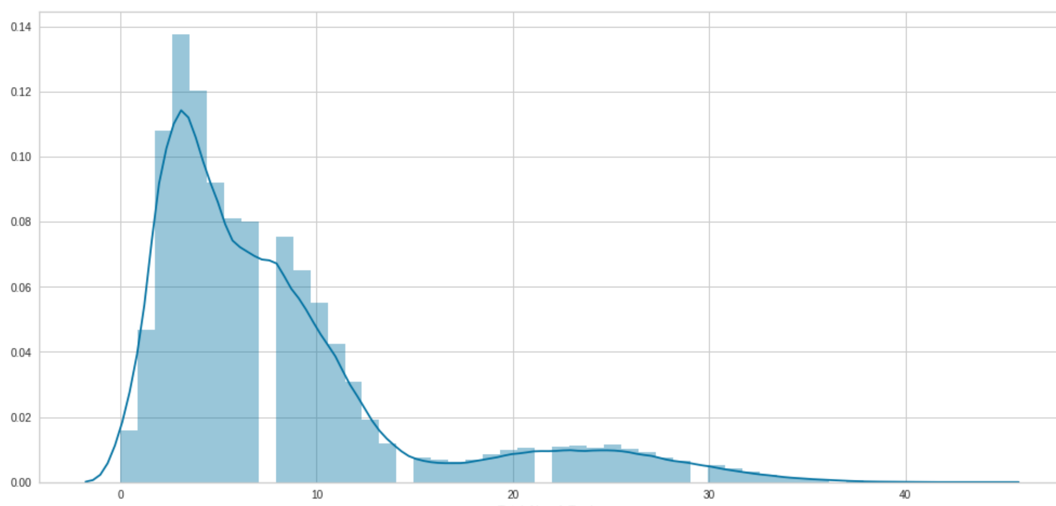
	bins	Defaulters	Count	Non-Defaulters
0	0	80 7 .O	20 5 B 0	T02 73. 0
1	1	155 .0	3B99	3744 0
2	2	3B2.0	79D7	7525.0
3	3	444.0	B979	B 535.0
4	4	3BO .0	7113	67330
5	5	362.0	4926	45640
6	6	247.0	36 15	3368.0
7	7	209.0	2992	2 7B3.0
8	8	14 1 .0	2345	22040
9	9	71 .0	1777	1706.0
10	10	B4.0	1 JOB	14 24. 0
11	11	5 3. 0	1 23 1	11 7B .0
1 2	1 2	40.0	936	B96.0
13	13	26.0	7B9	763.0
14	14	23.0	553	5300
15	15	12.0	360	34BO
16	16	6.0	2 12	206.0
1 T	17	6.0	97	91 .0
1B	1B	0.0	40	40.0
19	19	0.0	6	6.0

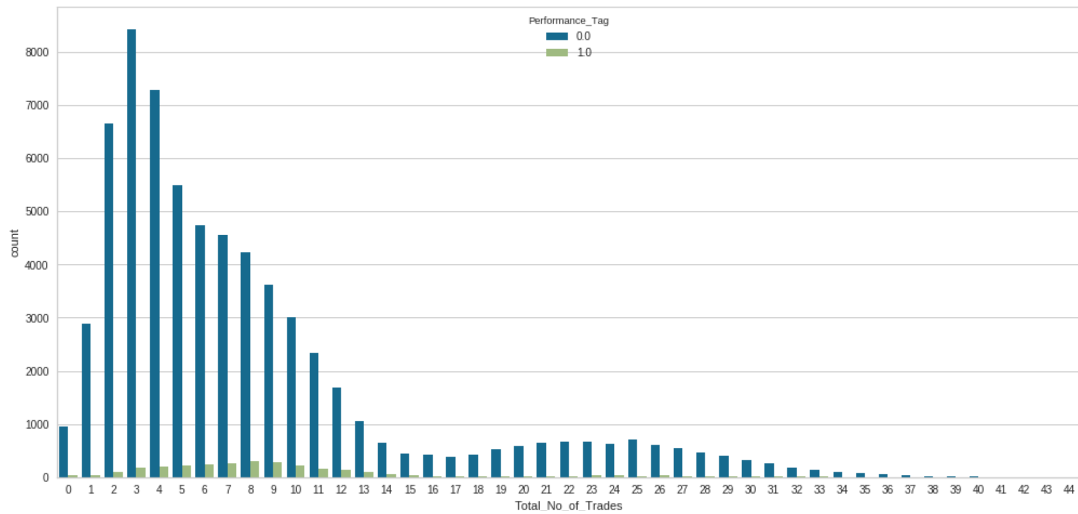
➢ **Presence_of_open_home_loan:**

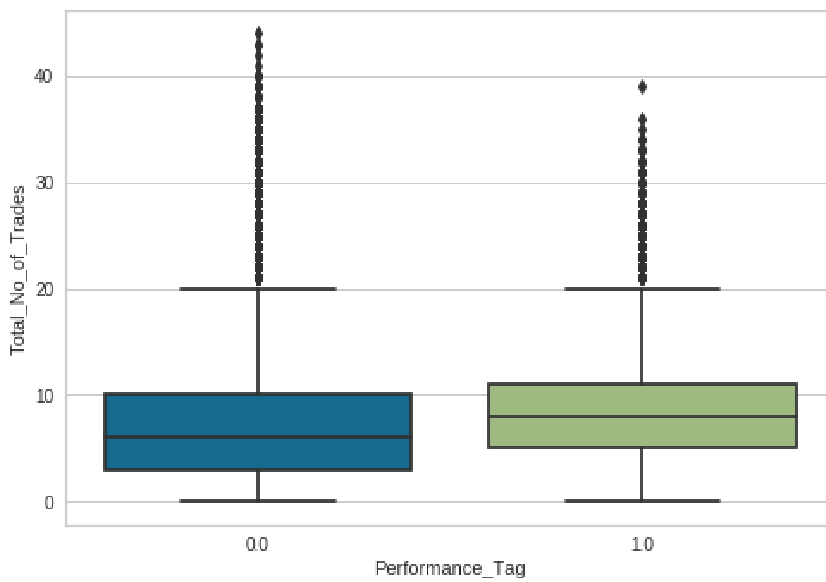


	bins	Defaulters	Count	Non-Defaulters
0	0.0	2333.0	51524	49191.0
1	1.0	607.0	18071	17464.0

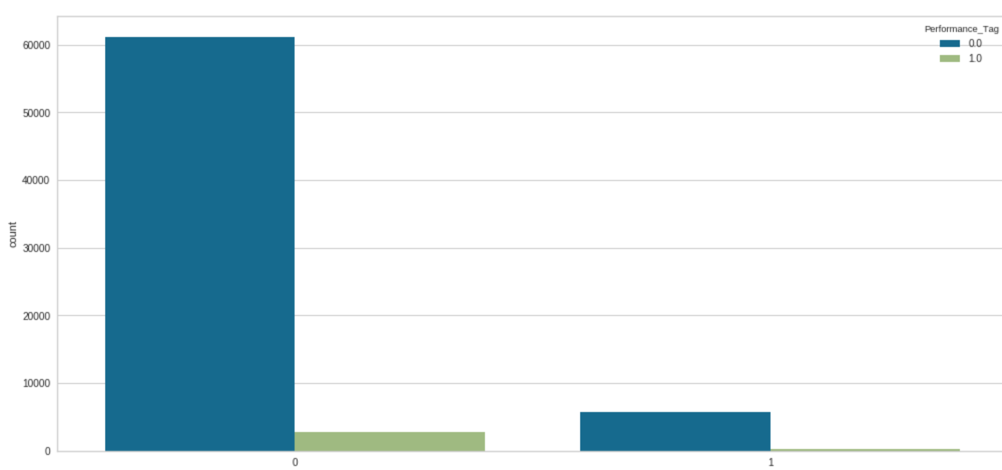
➢ **Total_No_of_Trades:**





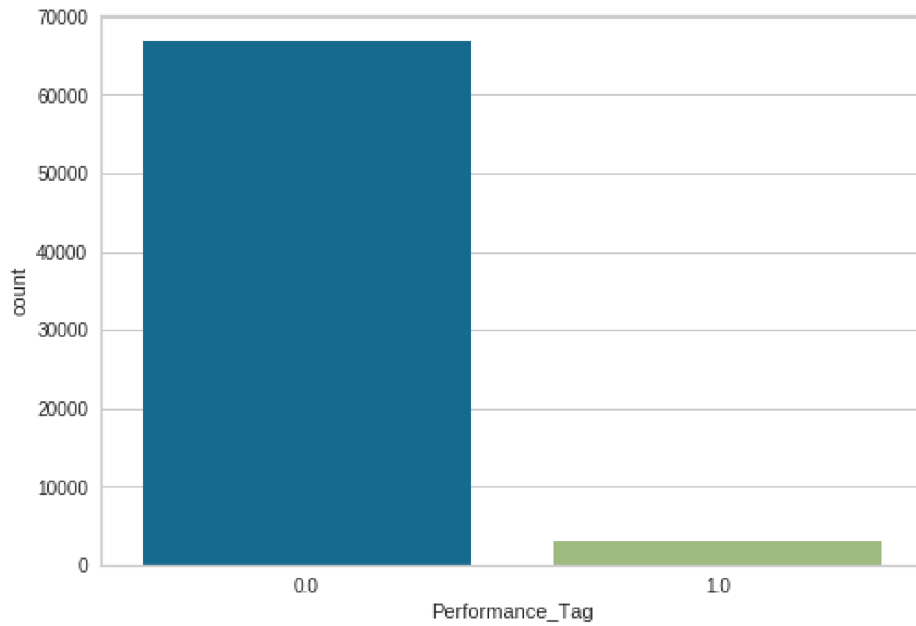


- **<u>Presence_of_open_auto_loan</u>**



	bins	Defaulters	Count	Non-Defaulters
0	0	2729.0	63938	61209.0
1	1	219.0	5929	5710.0

➢ **Performance_Tag:**



- There's high data imbalance between the two categories.

5) **Information Values of the Variables**

	Variable	IV
0	Application_ID	0.001487
0	Age	0.004169
0	Gender	0.000319
0	Marital_Status_at_the_time_of_application	0.000093
0	No_of_dependents	0.002652
0	Income	0.042842
0	Education	0.000764
0	Profession	0.002287
0	Type_of_residence	0.000920
0	No_of_months_in_current_residence	0.070893
0	No_of_months_in_current_company	0.022717

As we can see that only Income, No_of_months_in_current_residence and No_of_months_in_current_company are the features that can give us valuable information and can help in segregating Good from Bad customers. Hence we have build model only using these 3 Variables on Demographic dataset.

	Variable	IV
0	Application_ID	0.001487
0	No_of_times_90_DPD_or_worse_in_last_6_months	0.162992
0	No_of_times_60_DPD_or_worse_in_last_6_months	0.211549
0	No_of_times_30_DPD_or_worse_in_last_6_months	0.244473
0	No_of_times_90_DPD_or_worse_in_last_12_months	0.216024
0	No_of_times_60_DPD_or_worse_in_last_12_months	0.188546
0	No_of_times_30_DPD_or_worse_in_last_12_months	0.218904
0	Avgas_CC_Utilization_in_last_12_months	0.299389
0	No_of_trades_opened_in_last_6_months	0.187402
0	No_of_trades_opened_in_last_12_months	0.293711
0	No_of_PL_trades_opened_in_last_6_months	0.224320
0	No_of_PL_trades_opened_in_last_12_months	0.258756
0	No_of_Inquiries_in_last_6_months_excluding_hom...	0.113099
0	No_of_Inquiries_in_last_12_months_excluding_ho...	0.245292
0	Presence_of_open_home_loan	0.017010
0	Outstanding_Balance	0.245347
0	Total_No_of_Trades	0.232316
0	Presence_of_open_auto_loan	0.001662

Above we have IV of Credit Bureau dataset. It's quite evident that compared to Demographic dataset, credit bureau dataset carries a lot more information that can help us in achieving the task. Top 3 most important variables here are: Avgas_CC_utilization_in_last_12_months, No_of_trades_opened_in_last_12_months and No_of_PL_trades_opened_in_last_12_months.

6) Analysis so far and next steps:

- Most of the model we have built so far has a very low **precision** for class 1 (Default) category. Recall for both the classes is between 60-70% based on the algorithm.
- We are using SMOTEENN algorithm with worked better than SMOTE which are both up Sampling techniques but the former also does the job of clean up based on Nearest Neighbor algorithm.
- As per the business objective we need the model to be more accurate on identifying the False Negatives i.e, people who are likely to **Default (has a Performance Tag as 1 but predicted as 0).** This value should be very less when comparing and building each models and hence we should be looking at the Recall Metrix extensively while maintaining a balance on other metrics like Precision, Sensitivity and F1 score(Overall)…etc

- ➢ There will be a little bit of revenue loss for not considering Precision as the priority since we will incorrectly identify False Positives comparatively with the final model resulting in not issuing Credit cards to those customers in the first place. But as CredX is losing a lot of money due to customers going in Default status, then we shall build model keeping this our priority.
- ➢ As a next step, we need to build Scorecard based on the finalized model in order to predict the potential financial benefits for the company.
- ➢ For Application Scorecard implementation, we are given –

  "Build an application scorecard with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points."

  Hence, we can use below calculations to calculate the Scorecard:

  target_score = 400
  target_odds = 10
  pts_double_odds = 20
  factor = pts_double_odds / log10(2)
  offset = target_score - factor × log10(target_odds)
  scorecard['logit'] = ∑ (β×WoE) + α
  (where β – logistic regression coefficient and α – logistic regression intercept)
  Finally, scorecard['score'] = offset - factor × scorecard['logit']

7) Model Building Results along with KPI:

	Model	test_auc	train_sensitivity	test_sensitivity	delta_data_sensitivity	train_specificity	test_specificity	delta_data_specificity	delta_data_misclassifications
0	Logistic_Regression_Demographic	0.56	0.65	0.44	0.55	0.68	0.69	0	636
1	Decision_trees_Demographic	0.52	0.94	0.47	0.58	0.94	0.57	0	594
2	RandomForest_Demographic	0.56	0.84	0.4	0.62	0.86	0.71	0	535
3	Logistic_Regression_Combined	0.64	0.69	0.68	1	0.67	0.6	0	1
4	Decision_trees_Combined	0.52	0.89	0.1	0.27	0.96	0.95	0	1045
5	RandomForest_Combined	0.61	0.75	0.48	0.8	0.79	0.73	0	288
6	Gradient_Boosting_Combined	0.52	NA	NA	NA	NA	NA	NA	NA
7	Light_GBM_Combined	0.56	NA	NA	NA	NA	NA	NA	NA

Summary

- In the above Dataframe, NA means the data was not calculated if it performed poor in the initial model assessment.
- We can see that Logistic Regression on combined dataset performed best among the list of models with the highest AUC score of 64% and a very balanced sensitivity and specificity on both test and train dataset. Recall on the Delta dataset is also 100% with just 1 misclassification out of total 1425 applicants.
- Second best is Random Forest Model with an AUC score of 61%. A total of 288 applicants were misclassified by RF model on the delta dataset.

8) Model Evaluation Criteria:

- Optimum Sensitivity/Recall.
- Confusion matrix for each model.
- Sensitivity, specificity, AUC curve for Regression models.
- AUC-ROC curve for the Regression models using cut-off values for each model.
- Use of GridSearchCV/Random Search CV and plotting its results for all models.
- Gini-Index evaluation for Tree based models like decision tree and random forest.
- Within each model type evaluation using GridSerach based on recall values should be done to get models with optimized hyperparameters.
- For evaluation among models, the dataset for rejected applications (with performance tag missing), which were assumed as potentially defaulters should be considered for evaluations. Ideally, the output for all these applications should be defaulters or “1”.

9) Final Model Selection:

We selected Logistic regression as the final model for classification.

Reasons for Choosing this Model:

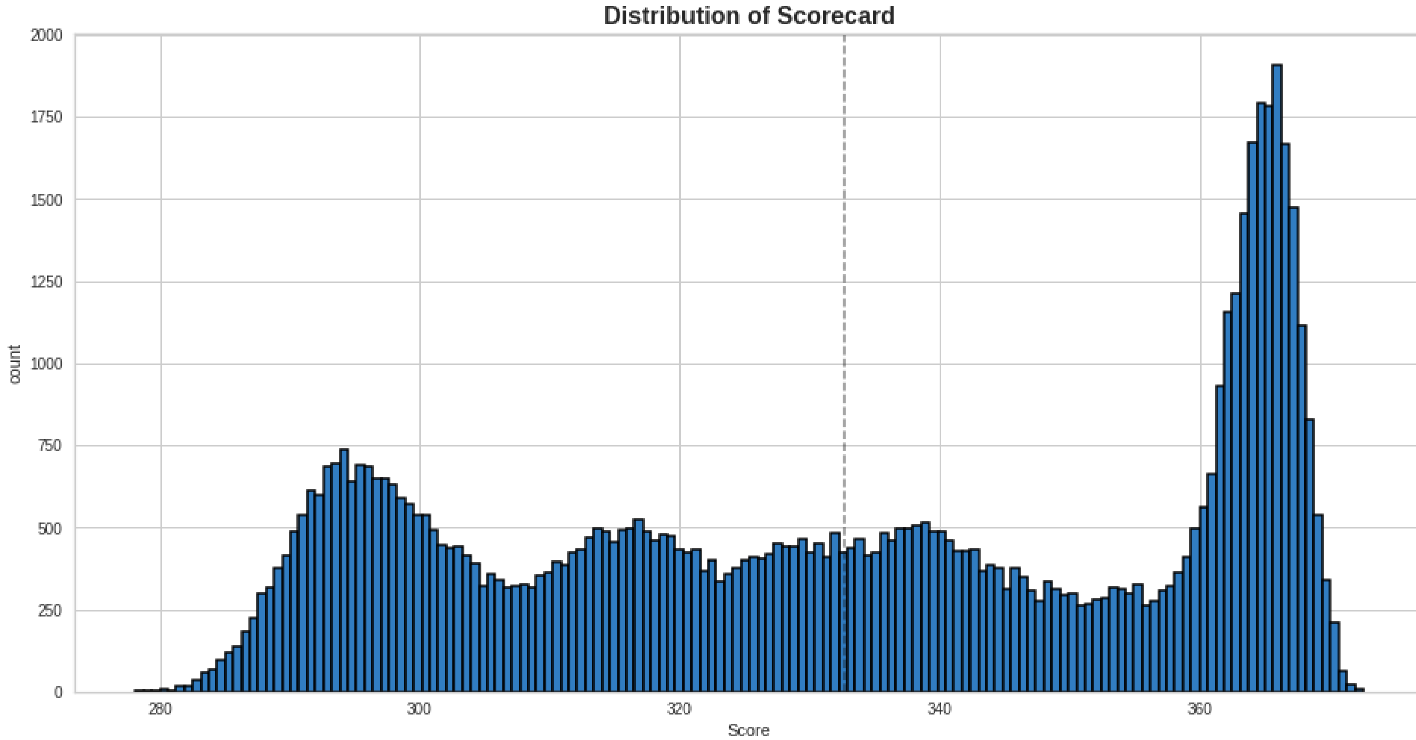
- The model gave good recall score on Test data (which is one of our objective).
- We got the best AUC score for this model comparatively with respect to the rest of the Models on both Test and Train sets.
- The model is not overfitting as there is very less difference in AUC score between train and test sets.
- The model was able to reject almost all the manually rejected applications (work like humans). Only one application classified as False Negative.
- The model is very stable. The use of WoE values makes it more robust. The WoE values are bound to show less variance making the model stable.
- The model is expected to have comparatively long life over others and is expected to have less modifications with time.

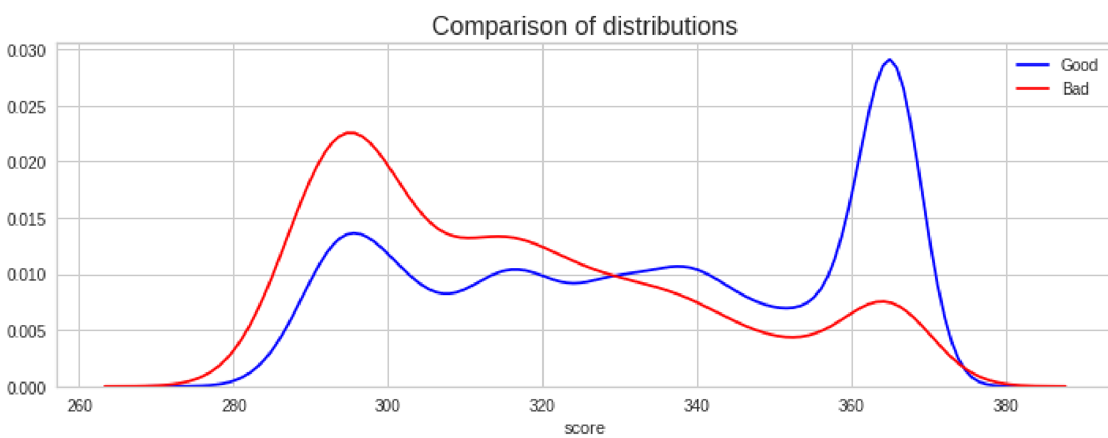
10) Application Scorecard:

Here are the formulae showing scorecard calculations:

```
score_df = new_applicants[coefficient_df.index].apply(lambda x:x*coefficient_df['Coef'],
score_df['logodds'] = score_df.sum(axis=1) + intercept
score_df['odds'] = np.exp(score_df['logodds'])
score_df['probs'] = score_df['odds'] / (score_df['odds'] + 1)
target = 400
odds = 10
doubleOdds = 20
factor = doubleOdds / np.log(2)
offset = target - (factor * np.log(odds))
score_df['score'] = offset - (factor * score_df['logodds'])
score_df['Performance_Tag'] = new_applicants['Performance_Tag']
score_df['score'] = round(score_df['score'], 2)
```

Distribution of Scorecard to find the optimal Cutoff:

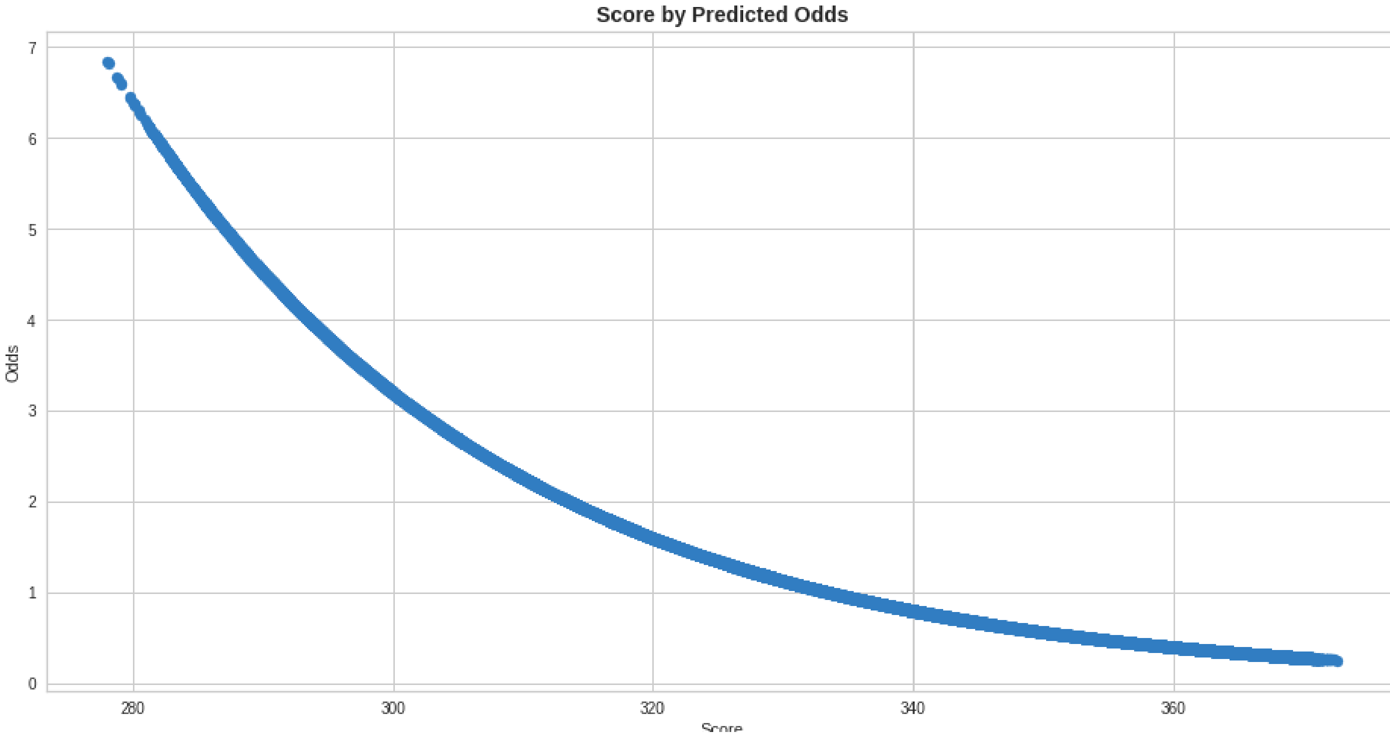




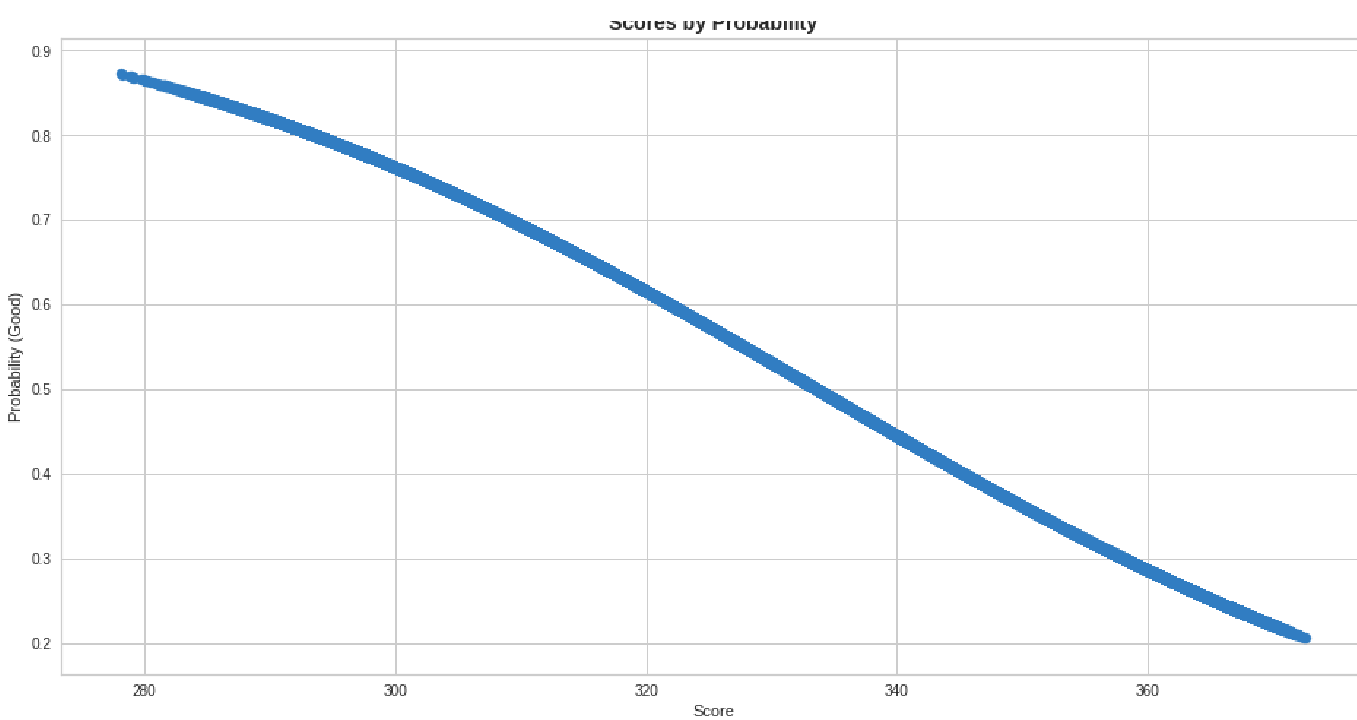

```
[ ] print("Final Cutoff Score is: ",score_masterdf['score'].mean())
```

```
Final Cutoff Score is:  332.6752581333119
```

Score of predicted Odds:



Scores by Probability:



11) Financial Benefits from model:

Financial Losses without model:

* A total of 71292 applications are available for issuing credit cards.
* 1425 applications were rejected by the bank.
* Number of people defaulted on their payments are 2948.

* Assumptions:
    * Consider cost of acquisitions to be approx 500 INR.
    * Approx Credit loss on principal from each applicants is 19500 INR.
* Total loss incurred would be (20000 * 2948) that is 58960000 INR.

Financial Benefits of Model:

* We have a recall rate of 68% hence we can straight away save 68% of the total losses that was incurred initially in decision making process without the use of Model; which is approx 40092800 INR
* Losses after using Model (58960000 - 40092800) is 18867200 INR, which is substantially lower than the initial losses.