CAPSTONE PROJECT
Final submission

Submitted by:

- Ravindra Vikram Singh
- Rituj Srivastava
- Ruchita Gupta
- Vipul Shrivastava

Business objective and problem statement

Problem Statement:

- CredX, a leading credit card provider, is experiencing huge credit losses.

Business Objective:

- To help CredX identify and 'acquire the right customers' and minimize losses.

Goal of Data analysis:

- To identify the right customers using predictive models.
- To determine the factors affecting credit risk using past demographic and credit bureau data of bank's applicants
- To create strategies to mitigate the acquisition risk and assess the financial benefit.
- To build an Application scorecard to identify good and bad customers

Data understanding & Cleaning- Demographic dataset

✓ Demographic dataset

- **71,295 application IDs in the dataset with 12 features**
- It includes Application ID, Age, Gender, Marital Status (at the time of application), No of dependents, Income, Education, Profession, Type of residence, No of months in current residence, No of months in current company, Performance Tag
- There are 1,425 null values in performance tag; these are the records of rejected candidates; removed these for model building, however, it will be useful for model validation at a later stage
- 3 duplicate Application IDs identified; decided to go with the record having Performance Tag as 1 or the latest record (deciding based on age)
- Removing Application ID column as it is all unique and not useful for analysis
- **Resultant data set has data for 69,867 application IDs with 11 attributes**

Data understanding & Cleaning- Credit Bureau dataset

✓ Credit Bureau dataset

- **71,295 application IDs in the dataset with 19 features**
- It includes Application ID, No of times 90 DPD (Days Past Due) or worse in last 6 months, No of times 60 DPD or worse in last 6 months, No of times 30 DPD or worse in last 6 months, No of times 90 DPD or worse in last 12 months, No of times 60 DPD or worse in last 12 months, No of times 30 DPD or worse in last 12 months, Avgas CC Utilization(Average utilization of credit card) in last 12 months, No of trades opened in last 6 months, No of trades opened in last 12 months, No of PL trades opened in last 6 months, No of PL trades opened in last 12 months, No of Inquiries in last 6 months (excluding home & auto loans), No of Inquiries in last 12 months (excluding home & auto loans), Presence of open home loan, Outstanding Balance, Total No of Trades, Presence of open auto loan, Performance Tag
- There are 1,425 null values in performance tag ; these are the records of rejected candidates; removed these for model building, however, it will be useful for model validation at a later stage
- 3 duplicate Application IDs identified; deleting the corresponding record which is deleted in Demographic data
- Removing Application ID column as it is all unique and not useful for analysis
- **Resultant data set has data for 69,867 application IDs with 18 attributes**

Methodology …

UpGrad

Problem Statement & Business objective

Demographic & Credit Bureau dataset

Performing Data quality checks

Exploratory data analysis

WOE transformation for treatment of missing values and outlier

Identifying important variables based on IV values

Removing highly correlated columns

Performing suitable treatment to handle imbalanced class

Merging datasets for Model building

Logistic Regression Model on Demographic WOE transformed data

Logistic Regression Model with Regularization and cross validation on WOE transformed merged data

Variable selection using Backward Elimination Method

Building Decision tress and Random Forests on variables finally selected

Model evaluation using sensitivity, specificity, ROC and KS-statistics

Model evaluation on the data for rejected candidates

Build Application Scorecard on selected Model for train, test and rejected data

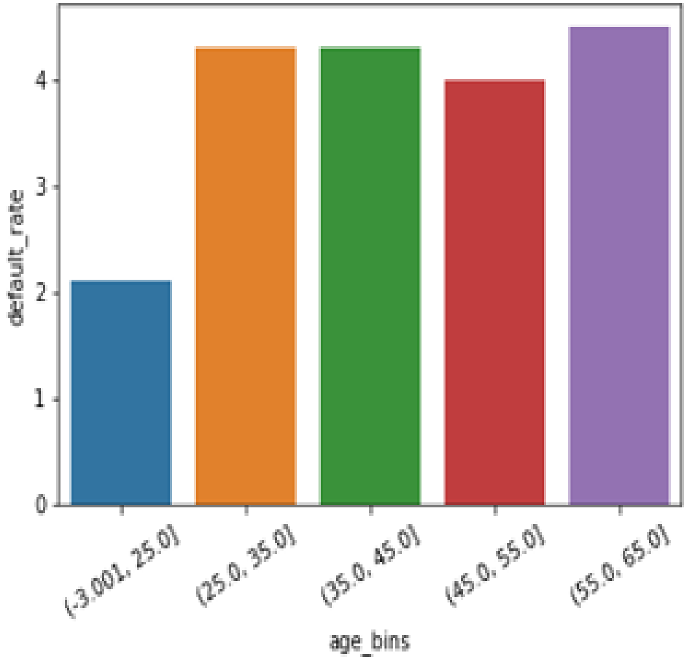
Financial Benefit Analysis and Recommendations

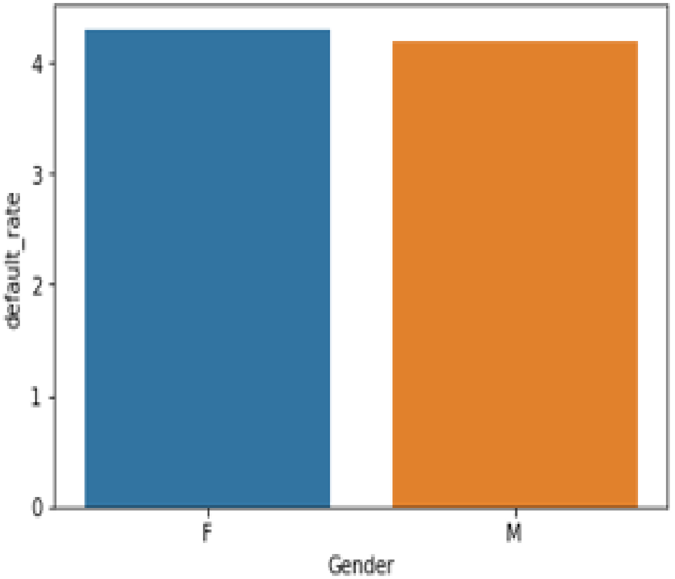
EDA: Key insights based on Demographic data (1/6)

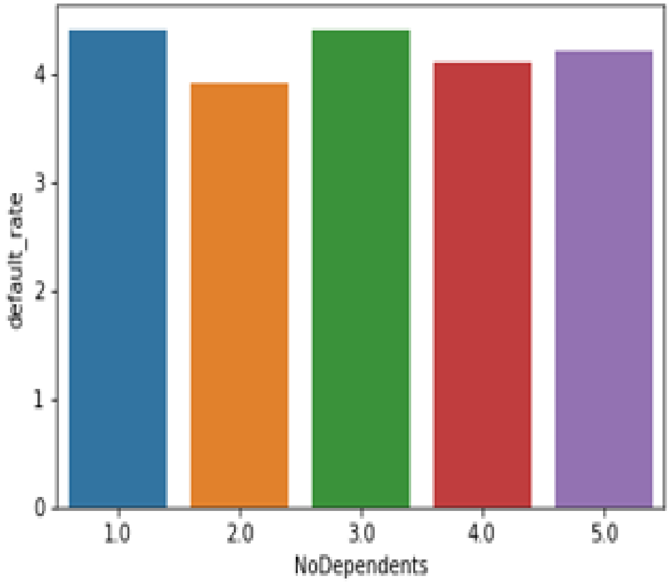
Default rate tends to increase with age: Customers in the higher age bracket likely to default

Females have slightly higher likelihood of default vs Males

No conclusive pattern of default rate emerging from no. of dependants







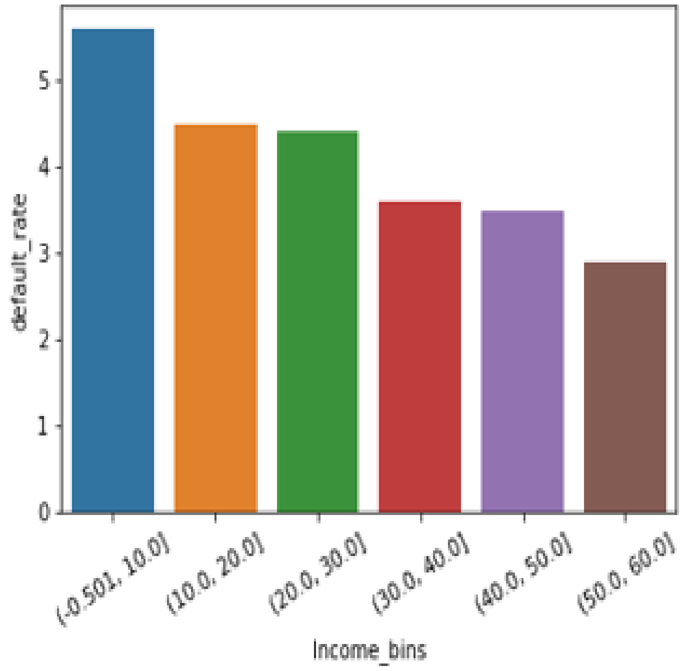
EDA: Key insights based on Demographic data (2/6)

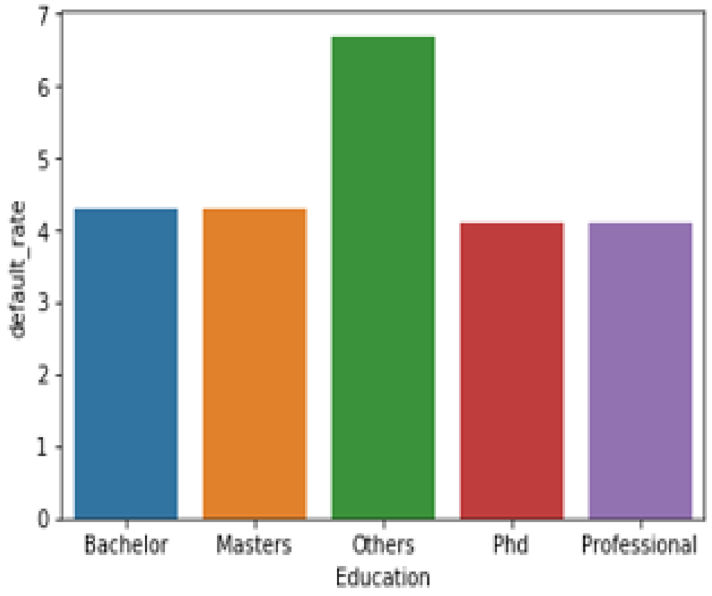
UpGrad

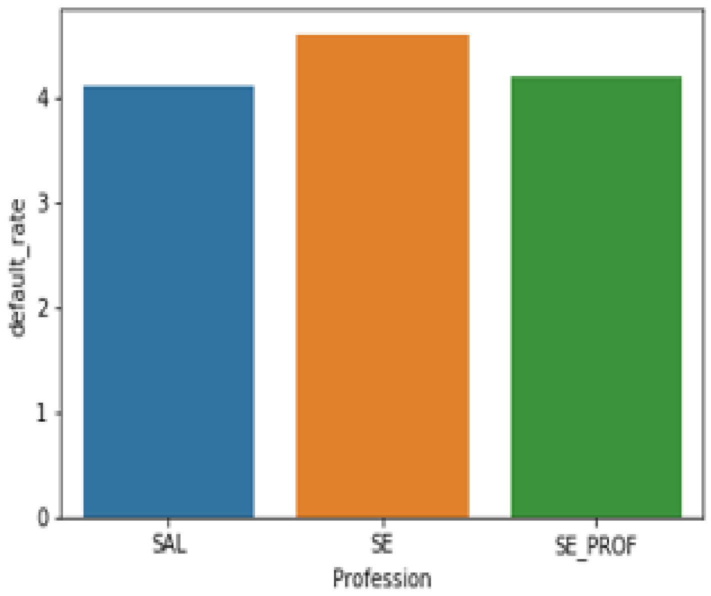
Default rate has an inverse correlation with income: it decreases with increase in income

Default rate high with “Other “ education category; however, count of such customers is very small to infer anything

Default rate slightly high for “SE” Profession





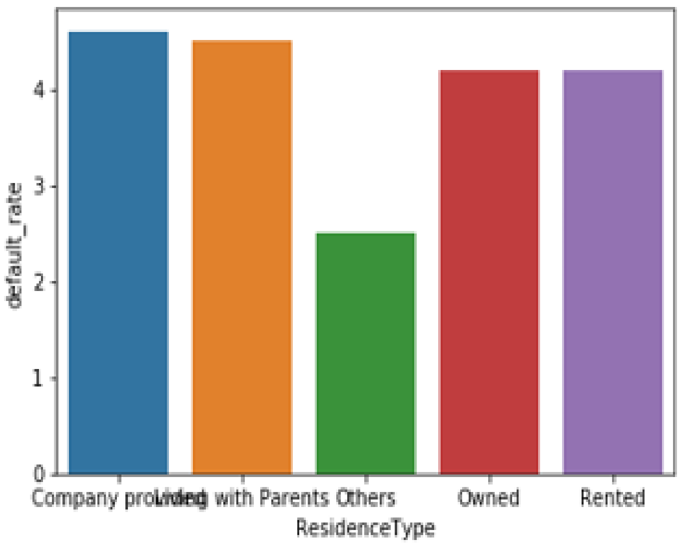


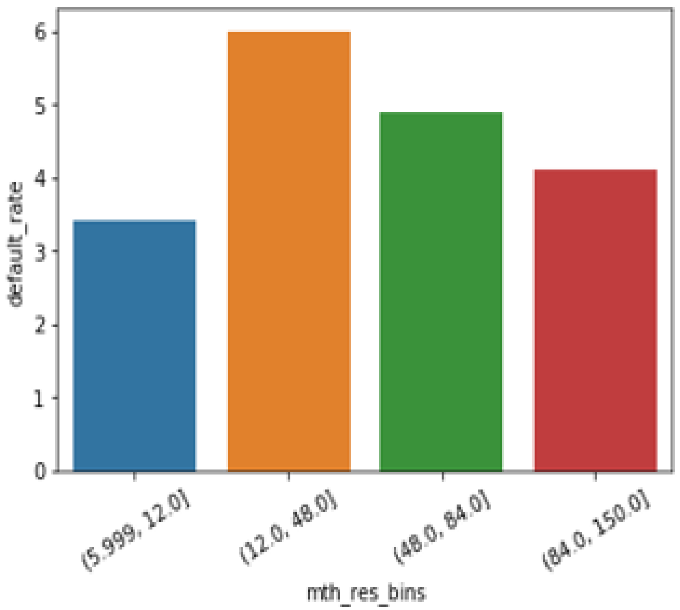
EDA: Key insights based on Demographic data (3/6)

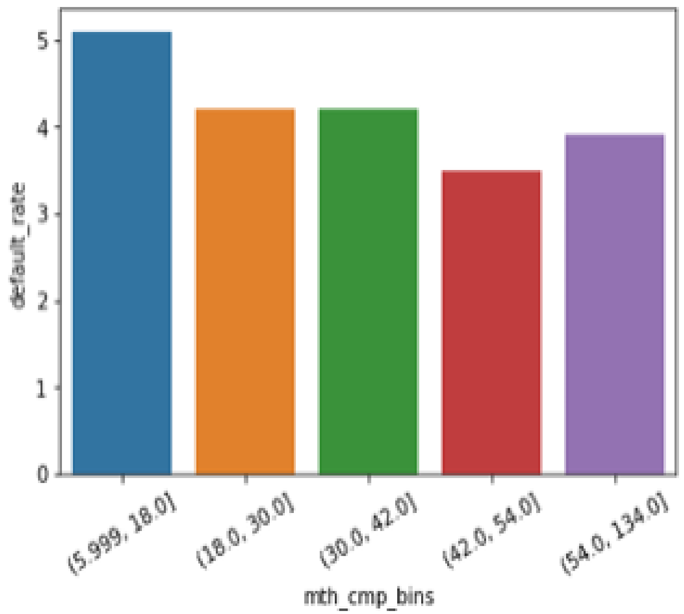
Default rate is less for Residence Type as Others

Default rate is higher for months of residence between 12-48

Default rate decreases as the customer's months in company increases with an exception for more than 5 yrs







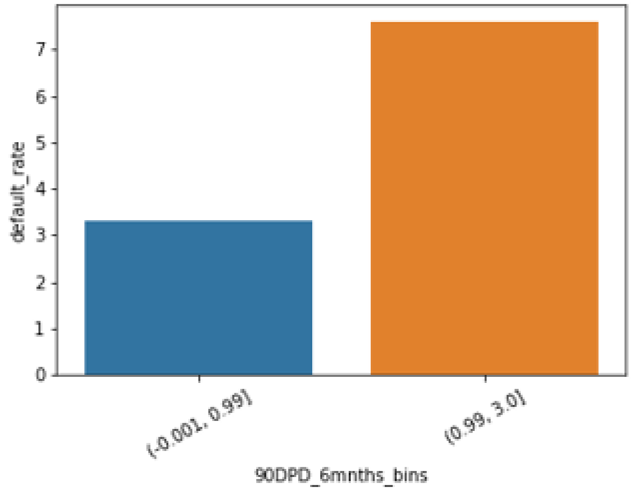
From the EDA on demographic data, Income, months of Residence and months in company look to be slightly important predictors.

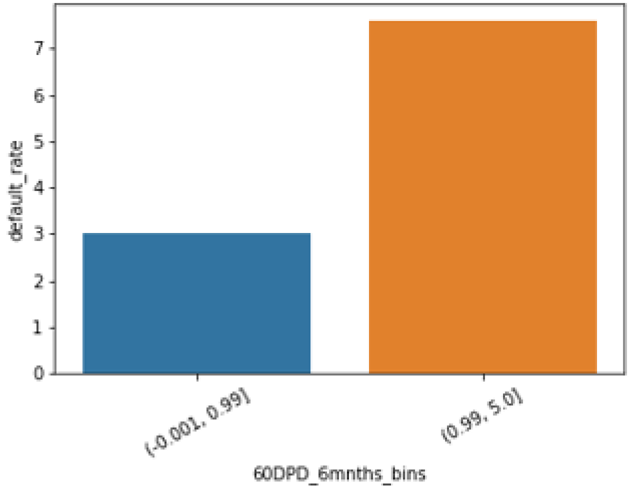
EDA: Key insights based on Credit Bureau data (4/6)

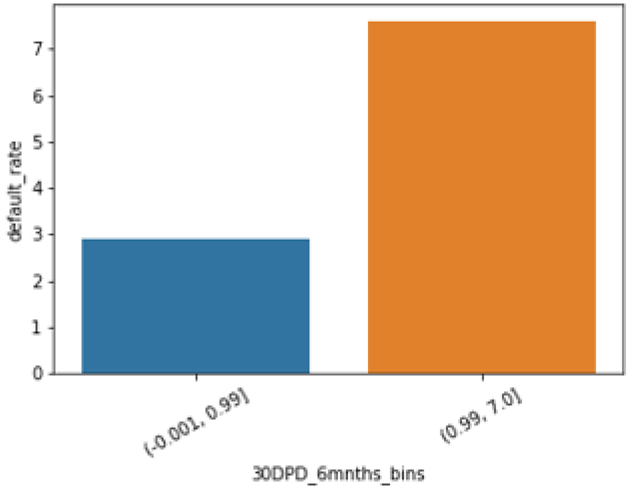
Default rate remarkably high when customer has > 1 90DPD in 6 and 12 months

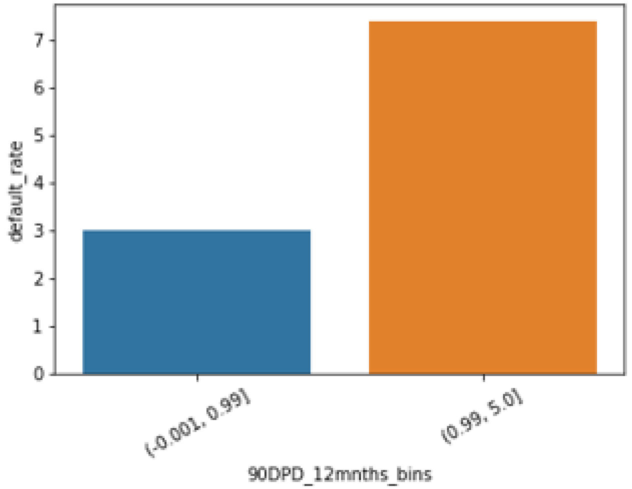
Default rate remarkably high when customer has > 1 60DPD in 6 and 12 months

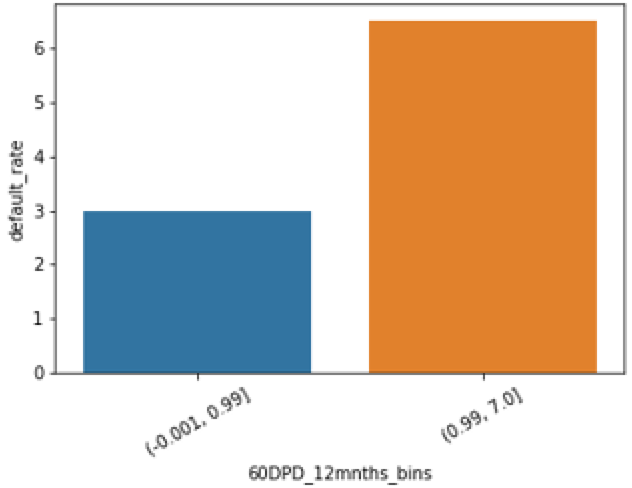
Default rate remarkably high when customer has > 1 30DPD in 6 and 12 months

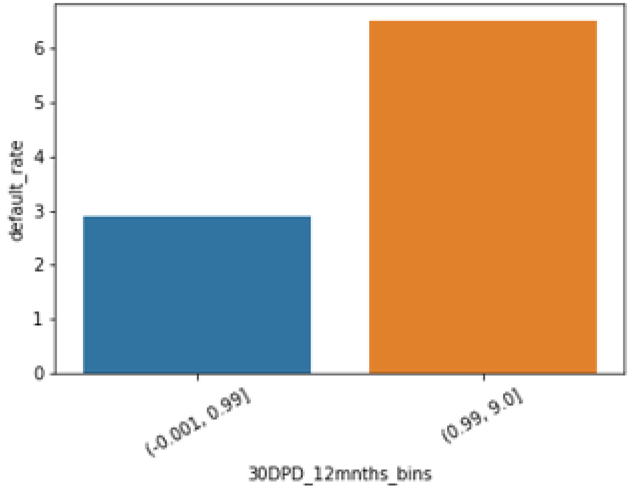




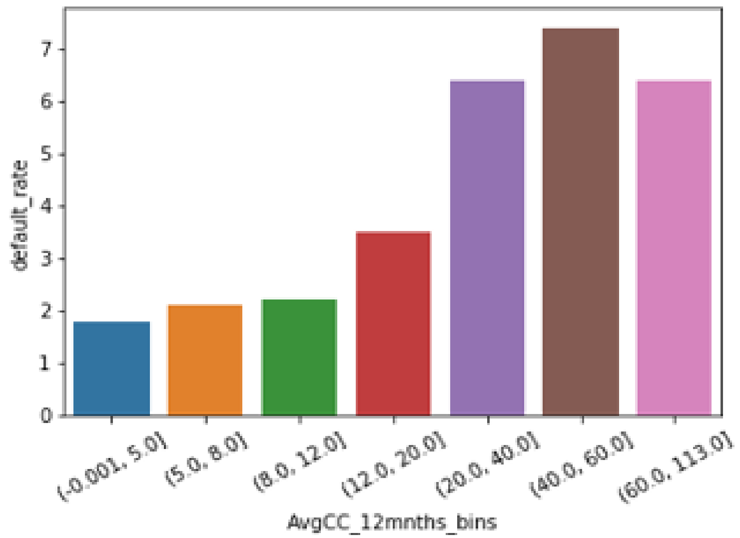




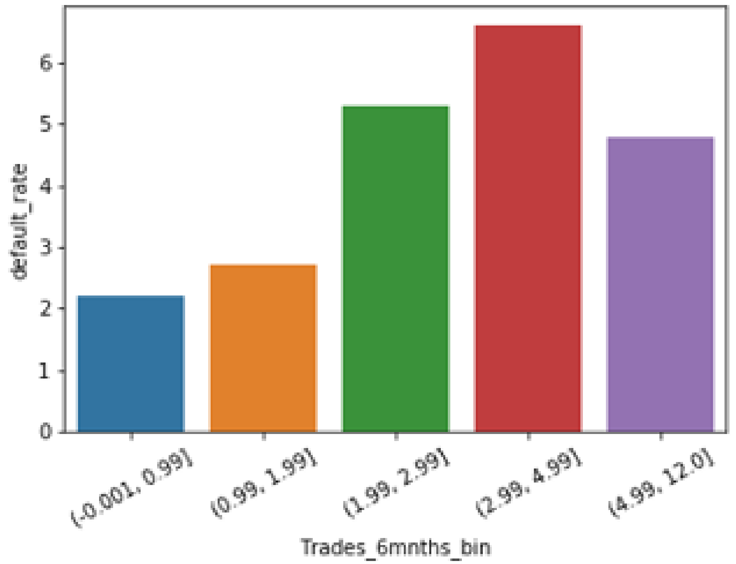




EDA: Key insights based on Credit Bureau data (5/6)

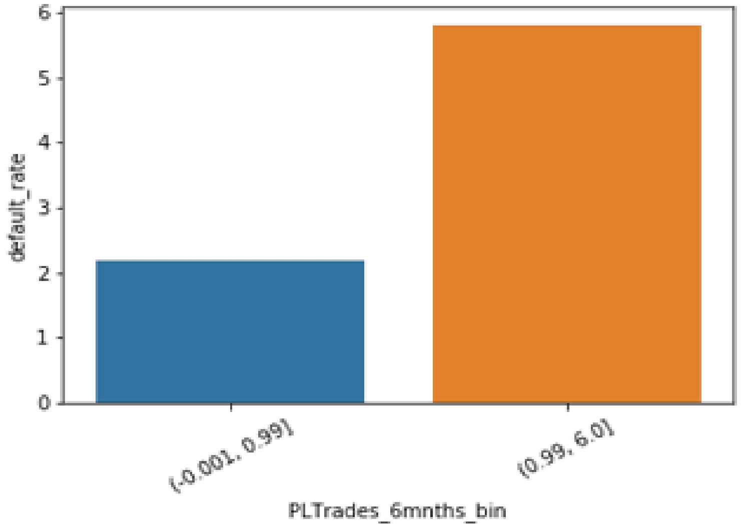


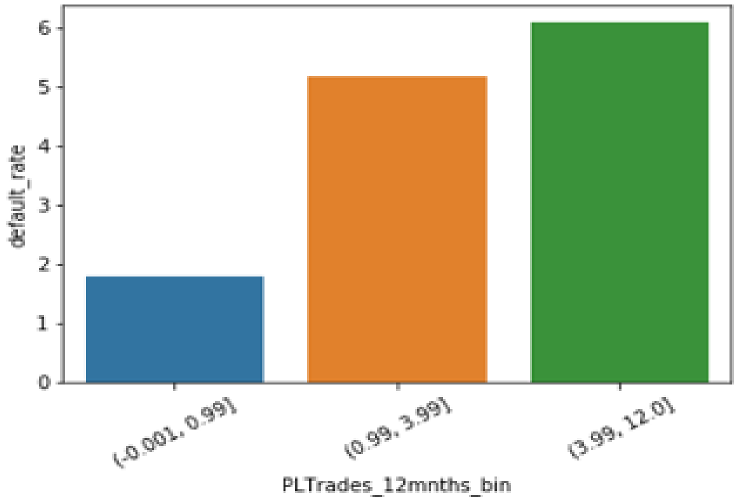
Default rate is increasing with increase in average Credit card utilization in last 12 months



Default rate remarkably high when the customer has 3 or 2 trades in the last 6 months

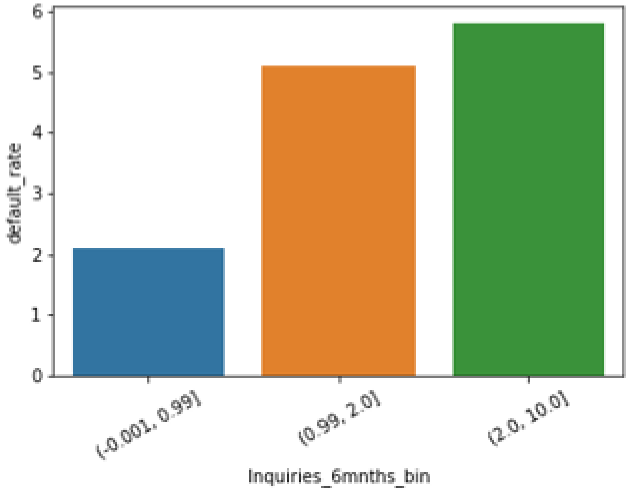
Default rate significantly high for more than 1 number of PL Trades in last 6 and 12 months

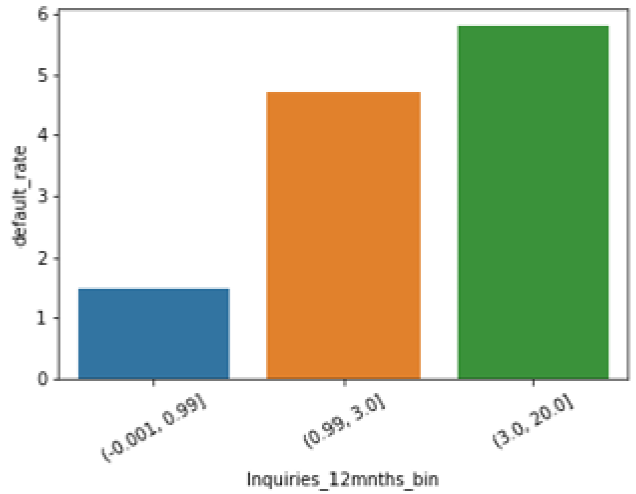




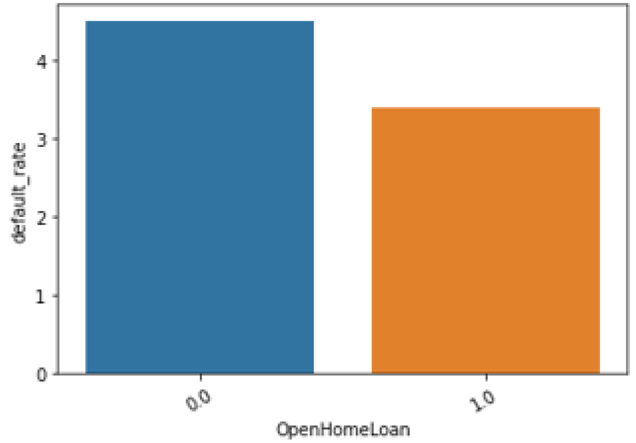
EDA: Key insights based on Credit Bureau data (6/6)

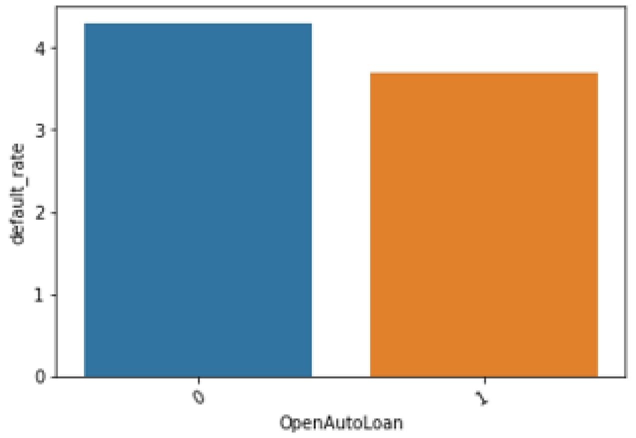
Default rate increases as the no. of Inquiries in the last 6/ 12 months increases



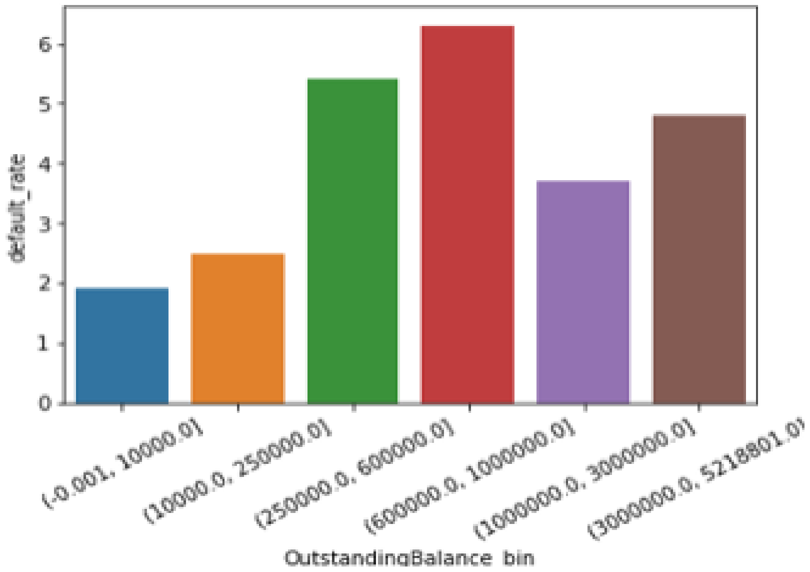


Customers who do not have an Open Home or Auto loan are more likely to default





Default rate increases with increase in the outstanding balance but falls if it is greater than 10000k



From EDA on credit Data, 90DPD in 6 and 12 months, 60 DPD in 6 and 12 months, 30 DPD in 6 and 12 months, Average card utilization, Trades and PL trades in 6 and 12 months, No. of Inquiries, outstanding balance and Open home or auto loan look to be strong predictors. But they can be correlated to each other.

WoE and IV analysis

- Weight of Evidence (WOE) and information value (IV) evolved from the logistic regression technique and have been in existence in credit scoring world for long.
- The weight of evidence tells the predictive power of an independent variable in relation to the dependent variable. Since it evolved from credit scoring world, it is generally described as a measure of the separation of good and bad customers.
- WOE can treat outliers and missing values through binning and WOE scores. Instead of using the raw values, we would be using WOE scores going forward.
- Since WOE Transformation handles categorical variable so there is no need for dummy variables.
- Information value(IV) is one of the most useful technique to select important variables in a predictive model. It helps to rank variables based on their importance.

IV < 0.02 :	predictor is not useful for modeling
0.02 < IV < 0.1:	predictor has only a weak relationship
0.1 < IV <0.3:	predictor has a medium strength relationship
0.3 < IV < 0.5:	predictor has a strong relationship

$$\text{WOE} = \ln\left(\frac{\text{\% of non-events}}{\text{\% of events}}\right)$$

$$\text{IV} = \sum (\text{\% of non-events} - \text{\% of events}) * \text{WOE}$$

WOE Analysis and Information value (IV)-Results

Demographic data

	VAR_NAME	IV
6	MnthsResidence	0.052
3	Income	0.038
5	MnthsCompany	0.012
8	Profession	0.002
1	Education	0.001
9	ResidenceType	0.001
0	Age	0.000
2	Gender	0.000
4	MaritalStatus	0.000
7	NoDependents	0.000

Credit Bureau data

	VAR_NAME	IV
6	AvgCC_12mnths	0.294
15	Trades_12mnths	0.258
7	Inquiries_12mnths	0.230
14	TotalTrades	0.190
0	30DPD_12mnths	0.189
12	PLTrades_12mnths	0.176
1	30DPD_6mnths	0.146
2	60DPD_12mnths	0.138
13	PLTrades_6mnths	0.125
4	90DPD_12mnths	0.096
16	Trades_6mnths	0.095
8	Inquiries_6mnths	0.092
3	60DPD_6mnths	0.090
5	90DPD_6mnths	0.030
11	OutstandingBalance	0.008
9	OpenAutoLoan	0.002
10	OpenHomeLoan	0.000

Demographic Data

- MnthsResidence and Income have **weak predictive power**
- Other variables are **not useful for model building**

Credit Bureau Data

- AvgCC_12mnths, Trades_12mnths, Inquiries_12mnths, TotalTrades, 30DPD_12mnths, PLTrades_12mnths, 30DPD_6mnths, 60DPD_12mnths and PLTrades_6mnths have **medium predictive power**
- 90DPD_12mnths, Trades_6mnths, Inquiries_6mnths, 60DPD_6mnths and 90DPD_6mnths have **weak predictive power**
- OutstandingBalance, OpenAutoLoan and OpenHomeLoan are **not useful for analysis**

Logistic Regression Model on Demographic data

Train data

Measure	Values
Sensitivity	58%
Specificity	56%
AUC	0.59

Test data

Measure	Values
Sensitivity	56%
Specificity	55%
AUC	0.57

- Demographic data has weak predictive powers as we have also seen with the Information Values

Logistic Regression Model on merged Demographic & Credit Bureau data with Cross Validation

- **Removed highly correlated variables along with application ID:**
  - WOE_30DPD_6mnths, WOE_Trades_6mnths, WOE_Trades_12mnths , ApplicationID
- **Model has been built on columns identified based on information values (IV)**
  - WOE_Income, WOE_MnthsResidence, WOE_AvgCC_12mnths, WOE_Inquiries_12mnths, WOE_TotalTrades, WOE_30DPD_12mnths, WOE_PLTrades_12mnths, WOE_60DPD_12mnths, WOE_PLTrades_6mnths, WOE_90DPD_12mnths, WOE_Inquiries_6mnths, WOE_60DPD_6mnths, WOE_90DPD_6mnths

Train data

Measure	Values
Sensitivity	70%
Specificity	56%
AUC	0.68

Test data

Measure	Values
Sensitivity	68%
Specificity	56%
AUC	0.67

- With similar statistics on train and test data, model is not overfitting

Logistic Regression Model on merged Demographic & Credit Bureau data using L1 Regularization and cross-validation

- **Removed highly correlated variables along with application ID:**
  - WOE_30DPD_6mnths, WOE_Trades_6mnths, WOE_Trades_12mnths, ApplicationID
- **Model has been built using L1 Regularization and deleting columns based on p-values**

Selected features

Features	VIF	Coeff
WOE_Income	1.10	-0.281624
WOE_MnthsCompany	1.03	-0.521061
WOE_Education	1.00	-0.345188
WOE_30DPD_12mnths	1.37	-0.467679
WOE_AvgCC_12mnths	1.78	-0.547920
WOE_Inquiries_12mnths	1.87	-0.469676
WOE_PLTrades_12mnths	1.86	-0.184068

Statistics

	Train data	Test data
Measure	**Values**	
Sensitivity	71%	69%
Specificity	56%	56%
AUC	0.68	0.67

- With similar statistics on train and test data, model is not overfitting

Decision tree Model on merged Demographic data & Credit Bureau data

Important Features

variables	importance_percentage
WOE_AvgCC_12mnths	61.447197
WOE_Inquiries_12mnths	14.610674
WOE_30DPD_12mnths	11.951371
WOE_MnthsResidence	4.839983
WOE_PLTrades_12mnths	2.555522
WOE_MnthsCompany	1.025535
WOE_Income	0.797972
WOE_Education	0.615288
WOE_ResidenceType	0.542158
WOE_Profession	0.465733
WOE_TotalTrades	0.393528
WOE_Inquiries_6mnths	0.317224
WOE_PLTrades_6mnths	0.226850
WOE_Age	0.210965
WOE_90DPD_6mnths	0.000000

Statistics

	Train data	Test data
Measure	**Values**	
Sensitivity	76%	72%
Specificity	53%	53%
AUC	0.69	0.67

Random Forest Model on merged Demographic data & Credit Bureau data

Top 15 Important Features

	variables	importance_percentage
	WOE_AvgCC_12mnths	31.906420
	WOE_Inquiries_12mnths	17.451290
	WOE_30DPD_12mnths	12.434277
	WOE_TotalTrades	8.896678
	WOE_PLTrades_12mnths	7.070911
	WOE_60DPD_12mnths	4.062714
	WOE_MnthsResidence	3.734437
	WOE_PLTrades_6mnths	2.088692
	WOE_Income	1.546294
	WOE_90DPD_12mnths	1.475284
	WOE_Inquiries_6mnths	1.465560
	WOE_MnthsCompany	1.218402
	WOE_60DPD_6mnths	1.163373
	WOE_Education	1.087622
	WOE_Age	0.805470

Statistics

	Train data	Test data
Measure	**Values**	
Sensitivity	74%	72%
Specificity	55%	54%
AUC	0.7	0.67

Final Model Selection

Metric Comparison Table

Models	Sensitivity		Specificity		ROC	
	Train	Test	Train	Test	Train	Test
Logistic Regression with IV selected Columns (13 columns)	70%	68%	57%	56%	0.68	0.67
Logistic Regression with L1 Regularization (7 columns)	71%	69%	56%	56%	0.68	0.67
Decision Tree	76%	72%	53%	53%	0.69	0.67
Random Forest	74%	72%	55%	54%	0.7	0.67

- Although Random Forest has better statistics than LR model, but since the LR model is less complex (only 7 features), we choose LR model. Moreover LR are linear models, and the logit-transformed prediction probability is a linear function of the predictor variable values which can be used to compute scores for the Application Scorecard Development

KS-Statistics for LR model

decile	total	default	cum_default	%cum_default	non_default	cum_non_default	%cum_non_default	Difference
1	2095	166	166	18.928164	1929	1929	9.604660	9.323504
2	2095	168	334	38.084379	1927	3856	19.199363	18.885016
3	2080	120	454	51.767389	1960	5816	28.958375	22.809014
4	2114	100	554	63.169897	2014	7830	38.986258	24.183640
5	2090	97	651	74.230331	1993	9823	48.909580	25.320751
6	2092	71	722	82.326112	2021	11844	58.972316	23.353795
7	2101	69	791	90.193843	2032	13876	69.089823	21.104020
8	2074	24	815	92.930445	2050	15926	79.296953	13.633492
9	2103	31	846	96.465222	2072	17998	89.613623	6.851600
10	2117	31	877	100.000000	2086	20084	100.000000	0.000000

- KS-Statistics for finally selected LR model on test data is 25.32 and is present in the 5^{th} decile.

Application Scorecard

- An application scorecard is built with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points.
- Using the LR model, we get the probability of default (bad) and non-default (good). The following formula is used to calculate the scores for each candidate :

slope=20/(ln(20)-ln(10)) = 28.85

Odds(good) = P(good)/P(bad)

LogOdds = ln(Odds(good))

Score = 400 + slope *(LogOdds – ln(10))

- After calculating the scores a cut-off score is calculated based on which it can be decided whether to grant or reject the credit card to applicants.

CUTOFF_SCORE = 400 + (slope * (log((1-x)/x) - log(10)))

where x is the Cutoff selected for probability of default in LR

Putting x=0.5, we get **CUTOFF_SCORE = 334**

ApplicationID	PGood	PBad	Predicted	Odds	LogOdds	Score
806528089	0.538450	0.461550	0.0	1.166612	0.154104	338.01
587334872	0.709384	0.290616	1.0	2.440968	0.892395	359.31
71610719	0.759219	0.240781	0.0	3.153149	1.148402	366.70
730740700	0.546366	0.453634	0.0	1.204423	0.186000	338.93
902449396	0.346327	0.653673	1.0	0.529818	-0.635222	315.23
526710823	0.584258	0.415742	0.0	1.405340	0.340279	343.38
429042181	0.752021	0.247979	1.0	3.032603	1.109421	365.57
529339062	0.737467	0.262533	0.0	2.809041	1.032843	363.36
814412959	0.453000	0.547000	0.0	0.828154	-0.188556	328.12
448532151	0.574384	0.425616	0.0	1.349534	0.299760	342.21
217649604	0.740015	0.259985	1.0	2.846377	1.046047	363.74
89399896	0.762464	0.237536	0.0	3.209890	1.166237	367.21
613414022	0.471826	0.528174	0.0	0.893316	-0.112815	330.31
239545054	0.410578	0.589422	1.0	0.696578	-0.361575	323.13
333988832	0.650092	0.349908	1.0	1.857893	0.619443	351.43

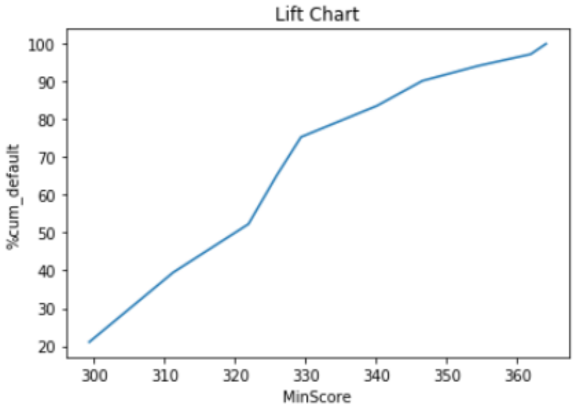
Lift Table and Gain Chart

- On Train

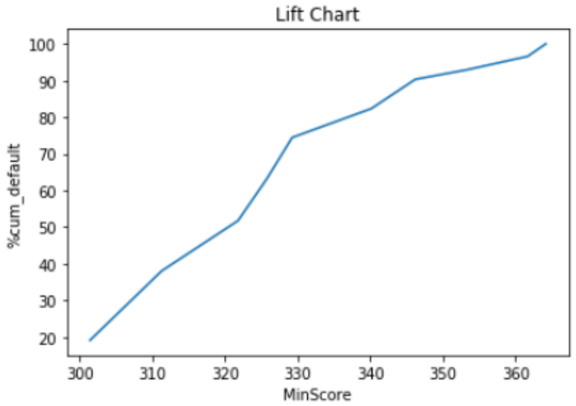
decile	total	default	MinScore	MaxScore	cum_default	cum_total	%cum_default	cumlift
1	4904	436	299.32	311.23	436	4904	21.052632	2.105263
2	4884	382	311.24	321.89	818	9788	39.497827	1.974891
3	4922	265	321.91	325.87	1083	14710	52.293578	1.743119
4	4866	265	325.89	329.34	1348	19576	65.089329	1.627233
5	4892	212	329.36	340.11	1560	24468	75.325930	1.506519
6	4888	171	340.12	346.57	1731	29356	83.582810	1.393047
7	4878	137	346.58	354.08	1868	34234	90.197972	1.288542
8	4931	82	354.48	361.93	1950	39165	94.157412	1.176968
9	4862	64	361.94	364.11	2014	44027	97.247706	1.080530
10	4879	57	364.12	367.78	2071	48906	100.000000	1.000000

- On Test

decile	total	default	MinScore	MaxScore	cum_default	cum_total	%cum_default	cumlift
1	2123	168	301.39	311.26	168	2123	19.156214	1.915621
2	2071	166	311.27	321.76	334	4194	38.084379	1.904219
3	2110	120	321.77	325.71	454	6304	51.767389	1.725580
4	2083	100	325.73	329.21	554	8387	63.169897	1.579247
5	2108	99	329.23	340.11	653	10495	74.458381	1.489168
6	2089	69	340.12	346.16	722	12584	82.326112	1.372102
7	2098	70	346.17	353.24	792	14682	90.307868	1.290112
8	2092	23	353.30	361.61	815	16774	92.930445	1.161631
9	2121	32	361.65	364.12	847	18895	96.579247	1.073103
10	2066	30	364.15	367.78	877	20961	100.000000	1.000000



On choosing the cut-off as 334, we are rejecting 50% of the customers and thus are able to reject around 75% of the defaulters. The Lift obtained is around 1.5 for both train and test data.



Financial Benefit of the Model

Assumptions :
Credit loss per defaulter = Rs 10,000/- Average revenue per customer = Rs. 500/-

Train Data

	Without Model	With Model
No. of Bad Customers	2071	511
No. of Good Customers	46835	23927
Credit Loss	2071 * 10000 = 20,710,000	511 * 10000 = 5,110,000
Total Revenue	46835 * 500 = 23,417,500	23927 * 500 = 11,963,500
Net Profit	23417500 - 20710000 = 2,707,500	11,963,500 – 5,110,000 = 6,853,500

- With the model, there has been an increase in the Net Profit from 2.7 million Rs to 6.85 million Rs
- Net financial gain = 4.1 million Rs
- Financial Gain % = 153%

Test Data

	Without Model	With Model
No. of Bad Customers	877	224
No. of Good Customers	20084	10242
Credit Loss	877 * 10000 = 8,770,000	224* 10000 = 2,240,000
Total Revenue	20084 * 500 = 10,042,000	10242 * 500 = 5,121,000
Net Profit	10,042,000 - 8,770,000 = 1,272,000	5,121,000 – 2,240,000 = 2,881,000

- With the model, there has been an increase in the Net Profit from 1.27 million Rs to 2.88 million Rs
- Net Financial gain = 1.6 million Rs
- Financial Gain % = 126.5%

Rejected Candidates

UpGrad

- On running the model on the rejected Candidates (with Performance Tag as N/A), around 99.4% candidates are predicted as Defaults. Hence our model is predicting quite accurately on the rejected candidates

- Built an Application Scorecard on them. The **minimum score assigned is 304** and **maximum score is 348**. On applying the **cutoff score = 334**, we get only 6 record as non-default. Hence, with our model and the chosen cut-off score, we would have been able to reject almost all the rejected candidates automatically.

Application Scorecard of Rejected Candidates

ApplicationID	PGood	PBad	Predicted	Odds	LogOdds	Score
906908303	0.315864	0.684136	1.0	0.461698	-0.772844	311.26
10990583	0.311952	0.688048	1.0	0.453387	-0.791009	310.74
589678446	0.303108	0.696892	1.0	0.434943	-0.832540	309.54
809411322	0.303764	0.696236	1.0	0.436294	-0.829438	309.63
150246616	0.321559	0.678441	1.0	0.473968	-0.746615	312.02
216681850	0.343597	0.656403	1.0	0.523455	-0.647303	314.88
413788459	0.353318	0.646682	1.0	0.546354	-0.604488	316.12
666004143	0.330262	0.669738	1.0	0.493122	-0.706999	313.16
505448697	0.281748	0.718252	1.0	0.392269	-0.935807	306.56
16819814	0.277576	0.722424	1.0	0.384229	-0.956515	305.96
597014646	0.349995	0.650005	1.0	0.538451	-0.619059	315.70
213641861	0.306974	0.693026	1.0	0.442947	-0.814304	310.07
937207017	0.298634	0.701366	1.0	0.425789	-0.853811	308.93
440239410	0.315913	0.684087	1.0	0.461803	-0.772618	311.27
640400319	0.347110	0.652890	1.0	0.531651	-0.631769	315.33

Conclusion

- **Logistic regression model is chosen** as the final Model with ROC score of 0.67
- Optimal **score cut-off value of 334** is derived to approve and reject the applications.
- Using this, we found out that our model is accurate in rejecting the candidate who may default in future. The % of **default customers without the model is 4.2%** whereas **with the model it is 2.1%**
- There is **Net Financial gain of 126% after using the model**.
- On the Rejected Candidates too, the model is predicting more than 99% of the customers as default.

Thank you