SIGMOID

SCALE | SPEED | INTELLIGENCE

Dhiraj Hasija

(Associate Data Scientist)

Mock Project

Project title: Prediction **of** Credit card eligibility and Credit limit determination.

Contents:

Problem Statement and understanding

Context

ABC Payment Bank, a mobile only bank provides digital credit card within an hour to eligible users with a credit limit from $1000 to $8000. However, it is challenging to decide whom to approve the card and credit limit to be given.

Objectives

1. Determine the eligibility of users for approval
2. Determine the credit limit for every customer.

Data

Previous customer experience data provided by the bank + CIBIL Data.

Domain Knowledge:

1. Gross annual Income
2. Work history
3. Education history
4. Credit history
   - Number of credit accounts
   - Length of credit history
   - Type of credit

Given Data:

1. Education
2. Occupation
3. Work class
4. Date of birth
5. Capital gain
6. Capital loss
7. Address
8. Hours per week
9. Marital status
10. Email
11. Inquiry purp code
12. Asset code
13. Asset class cd
14. Portfolio type
15. Institute type
16. Account type

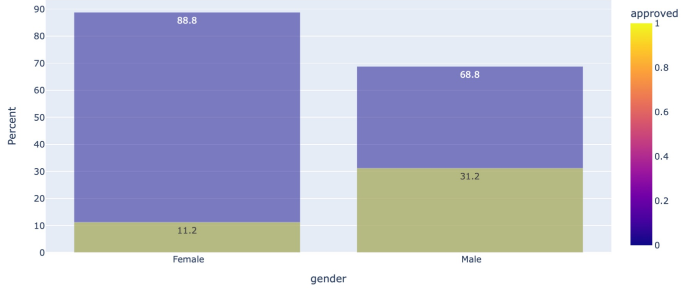
Key Insights from EDA

Categorically correlated features

1. Marital Status: 0.45
2. Education level: 0.36
3. Occupation: 0.35
4. Age group: 0.31
5. Hours group: 0.27
6. Inquiry purpose code: 0.21

Age	**Male**	**Female**
0-25	13%	22%
35-55	47%	39%

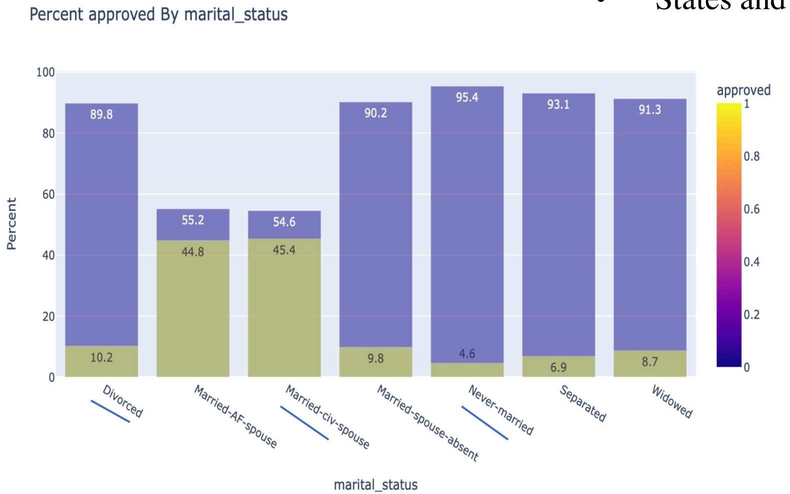
Age	**0-30 hrs**	**35-50 hrs**
0-25	41%	47%
35-55	7%	79%

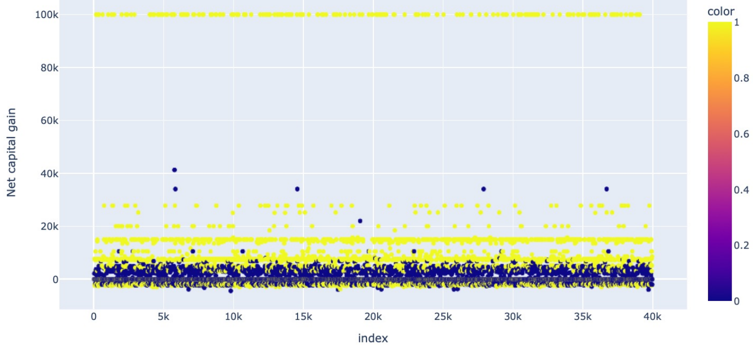


- Education level ⇔ Occupation
- Approval rate varies exponentially with education and occupation group
- States and Zipcode => America.

Highly correlated continuous Features:

1. Net income (occ, age, edu): 0.47
2. Age : 0.27
3. Hours per week: 0.23
4. Net capital gain: 0.21





Key Steps in feature engineering:

1. Adding incomes:

Incomes based on …[1]

Age	0.36
Education	0.37
Occupation	0.39
Relationship	0.34
Marital status	0.40
Workclass	0.23

Net income : 0.47

2. Net capital gain:

Capital gain – Capital loss

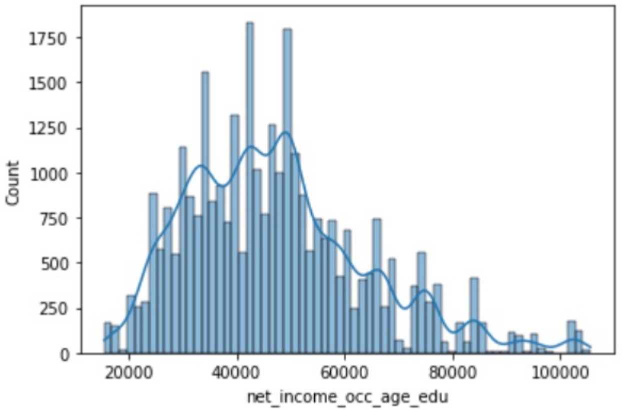
3. Education group:

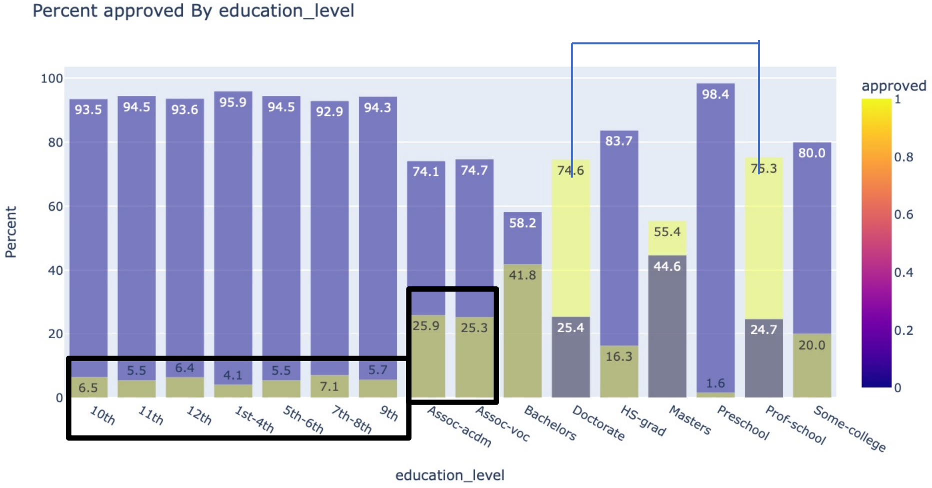
4. Age from DOB

5. Converted CIBIL data features to categories

*** Other Feature like state, Zipcode, category columns were Ignored due to lower correlation**

Missing value imputation : Mode





Type of encoding

Target encoding:

Account type, education group, asset code, inquiry purpose code, institute type, occupation, relationship, marital status, work class

1. Lesser number of features
2. Improves correlation between independent and dependent variables with its direct relationship
3. Faster learning
4. Can handle unknown values easily

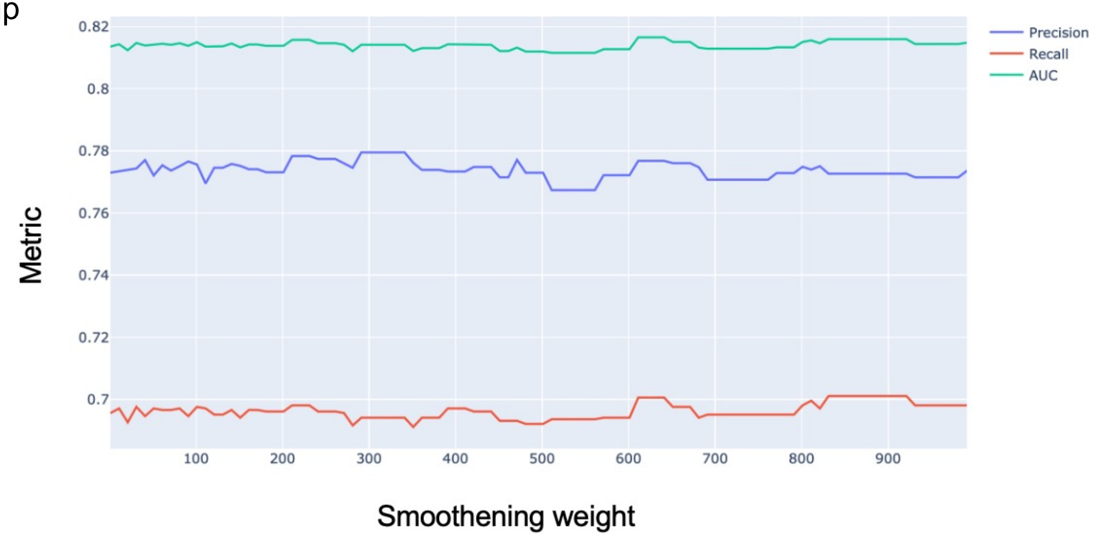
One hot encoding:

1. Large number of features
2. Slows the model
3. Multicollinearity

Label encoding :

Gender

1. Binary class

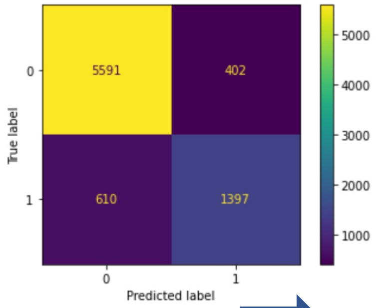


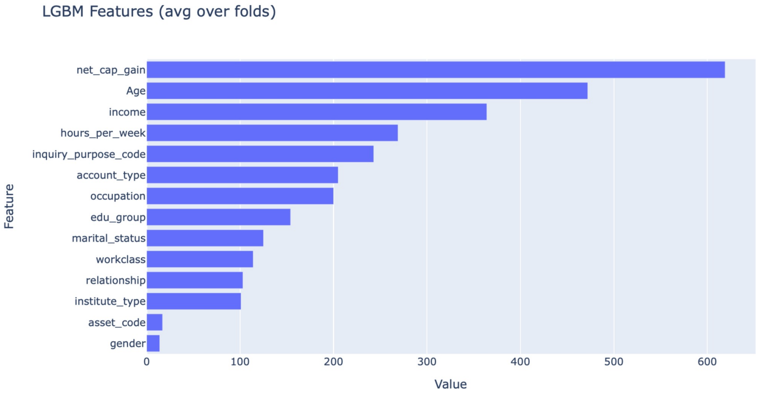
* Other columns like income, Age, hours per week were added directly

Best models and results:

Light GBM

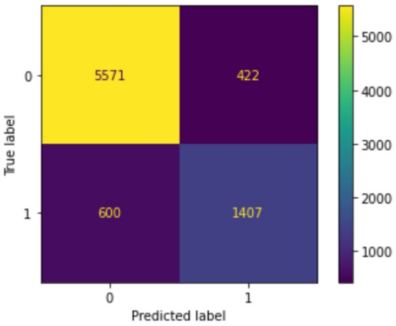
Metric	Value
Precision	77.7
Recall	69.6
Auc_roc score	81.6

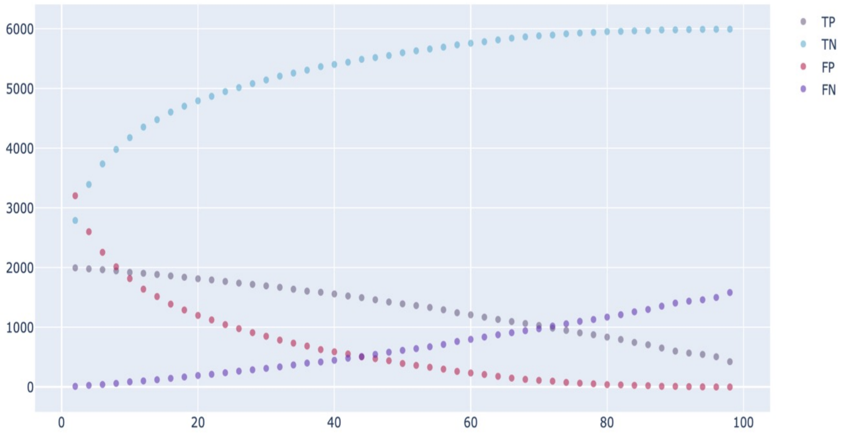




XG Boost

Metric	Value
Precision	76.9
Recall	70.1
Auc_roc score	81.4





Threshold = 0.42

Credit limit strategy:

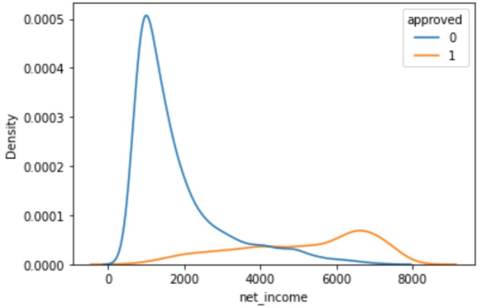
Credit

Credibility → Probability score from the model $\mathbf{p}_{\mathbf{score}}$

Credit capacity → Income score $\mathbf{i}_{\mathbf{score}}$ = Net_income ~ (0, 1)

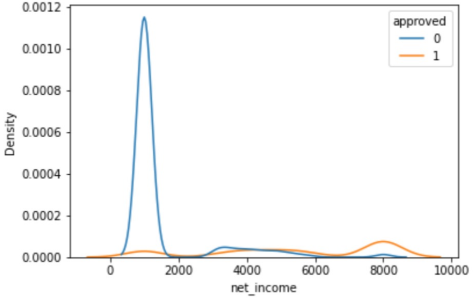
$$Net_{income} = w_1 * i_{age} + w_2 * i_{occ} + w_3 * i_{ed} + w_4 * i_{wcls} + w_5 * i_{rel} + w_6 * i_{msts} + w_7 * cap\_gain$$

Credit score = (0.65*$\mathbf{p}_{\mathbf{score}}$ + 0.35 * $\mathbf{i}_{\mathbf{score}}$) *8000



To increase the robustness and the revenue

If credit < $3000 => credit = $1000
If credit > $5800 => credit = $8000
If credit in (3000, 5800) => rounded to nearest 500



Tuned weights from feature importance

$\mathbf{w_1}$	10
$\mathbf{w_2}$	540
$\mathbf{w_3}$	200
$\mathbf{w_4}$	70
$\mathbf{w_5}$	50
$\mathbf{w_6}$	110
$\mathbf{w_7}$	80

Best numbers from the strategy

Revenue
$ 5.1M

51% of total credit

Actual loss
$ 1.42M

13% of total credit

Pseudo loss
$ 0.55M

5% of total credit

For validation set

Parameter	Value	% of total credit
Revenue	$ 9.9M	53.3%
Actual loss	$ 2.9M	15.9%
Pseudo loss	$ 0.39M	2.1%

Error for the model (%): 13 %

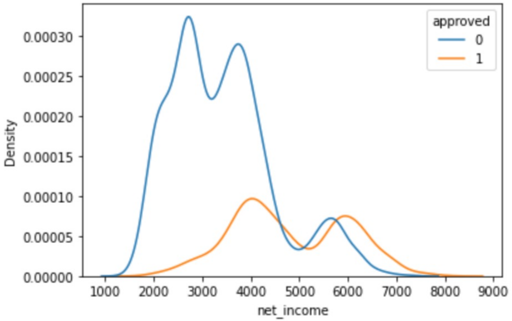
** Extra error in the credit is caused by the minimum limit of $1000

Appendix

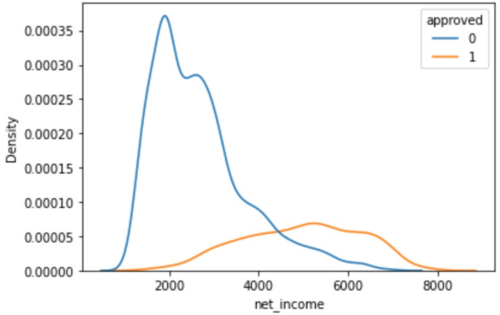
Link for income data

[1] https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-people.html

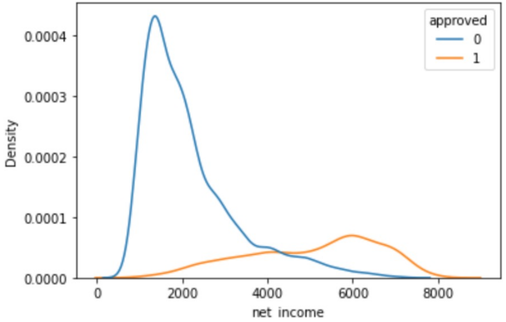
Variation of amount of credibility and credit capacity i.e. **a**



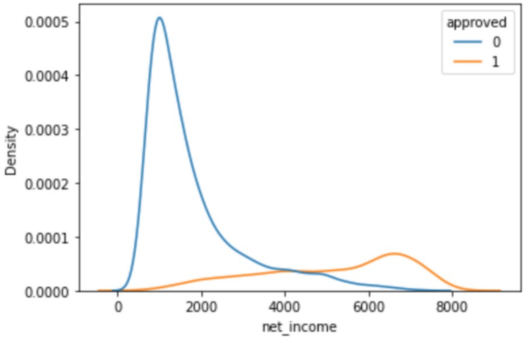
Pure salary, a = 1



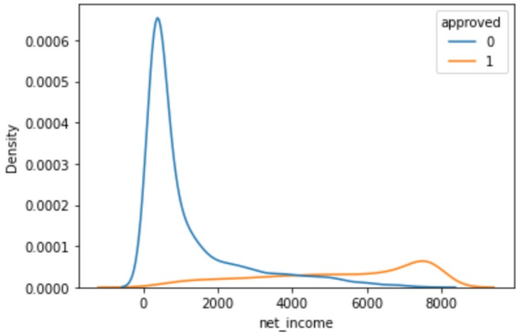
70% Salary,, a = 0.7



50% Salary, a = 0.5

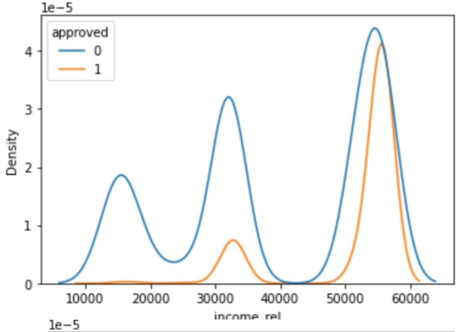


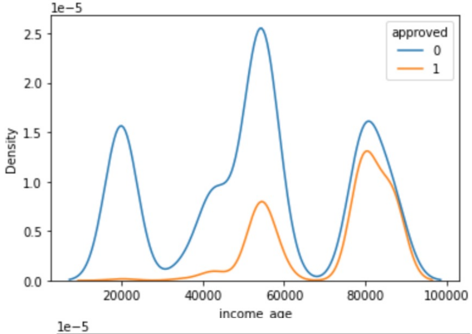
35% Salary, a=0.35

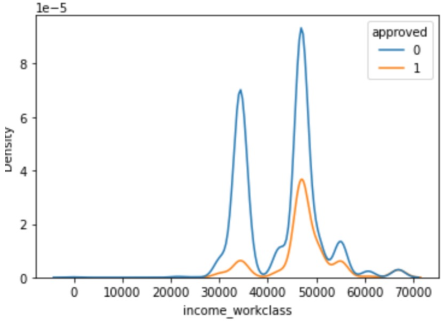


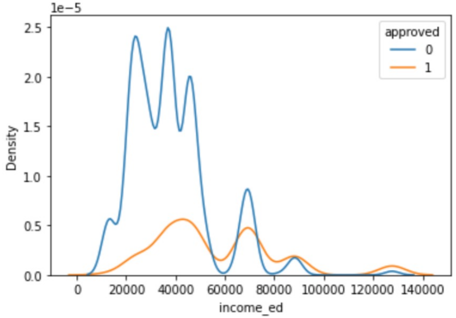
Pure prob, a = 0

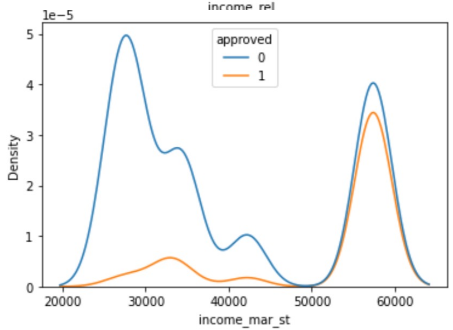
DISTRIBUTION OF INCOME WITH RESPECT TO DIFFERENT FEATURES

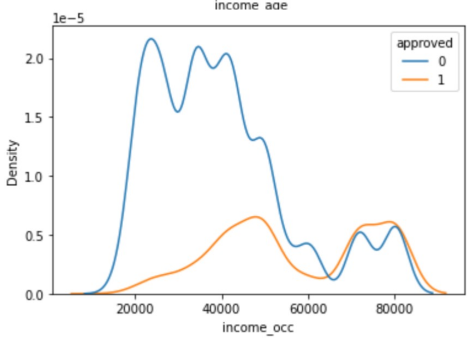




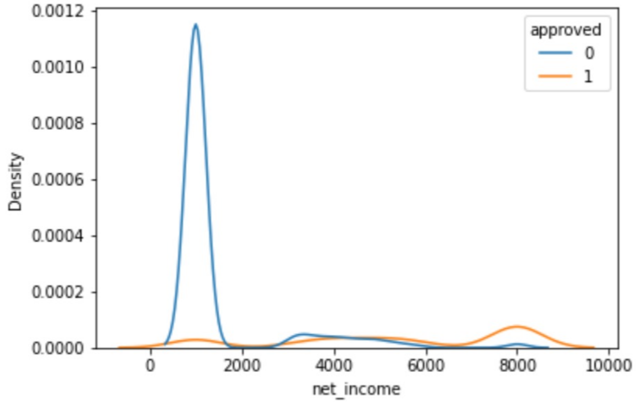




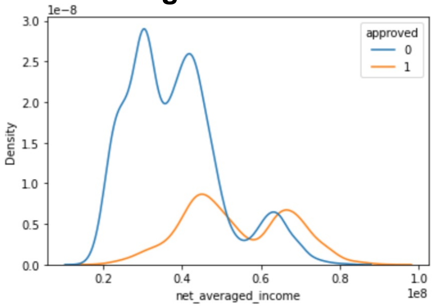




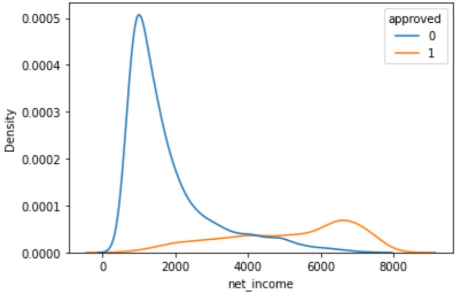
Distribution of credit



Distribution of Normalized weighted income

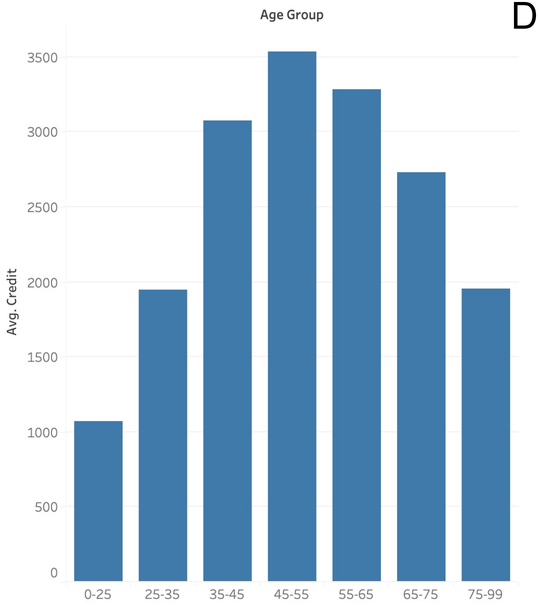


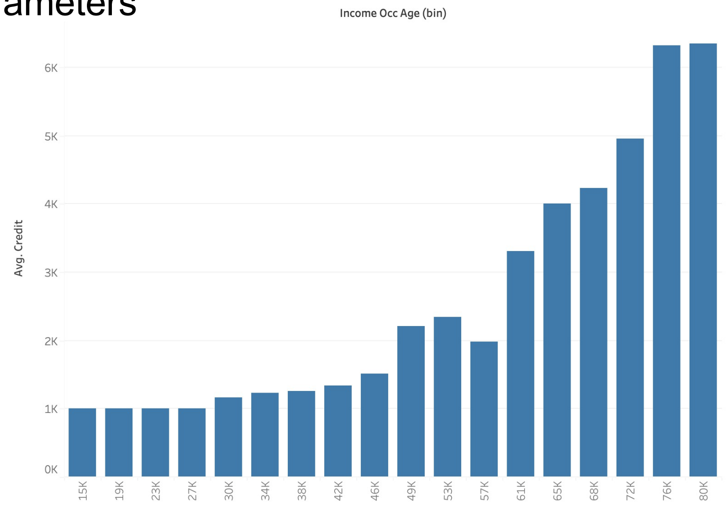
Distribution of income score

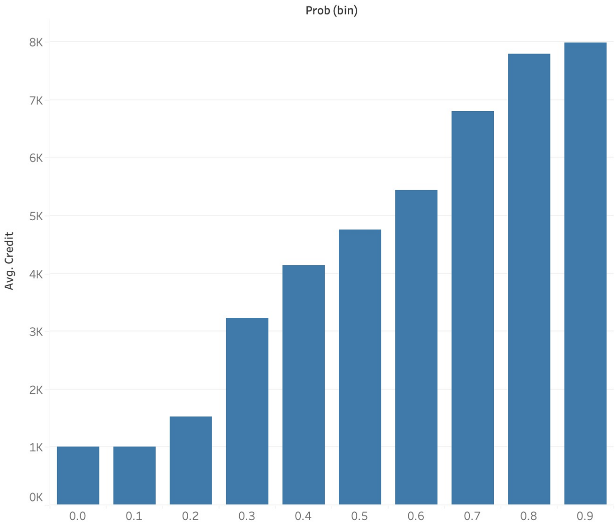


**** Income score is weighted income with probability**

Distribution of credit with different parameters







STEPS IN THE PROJECT

1. Understanding of Data, various features and their categories
2. Gained domain knowledge
3. Performed UVA and checked where the data cleaning is required
4. Performed BVA and MVA to determine the relationship between features
5. Performed feature engineering iterations for categorical features
    a. Tried combining categories
    b. Creating separate columns for categories with very high or low approval rate
    c. Combining features to create a new features.
6. Performed feature engineering iterations for continuous features
    a. Binning and converting to categorical
    b. Mathematical transformation to increase spacing between approved and non approved
    c. Combining multiple features.
    d. Adding new feature, salary and checking its correlation
7. Tried different encodings and tuned them.
8. Tried, Random forest, Logistic regression, XG Boost, Light GBM, Catboost
9. Tried combinations of different categorical and continuous features for making the model.
10. Hyper parameter tuning
11. Selected threshold.
12. Different credit limit strategies
    Based on approval.
    Based on threshold, two threshold
    Based on probability + feature
    Based on income
    Based on income and probability

13. Creating the prediction pipeline.

Thank you