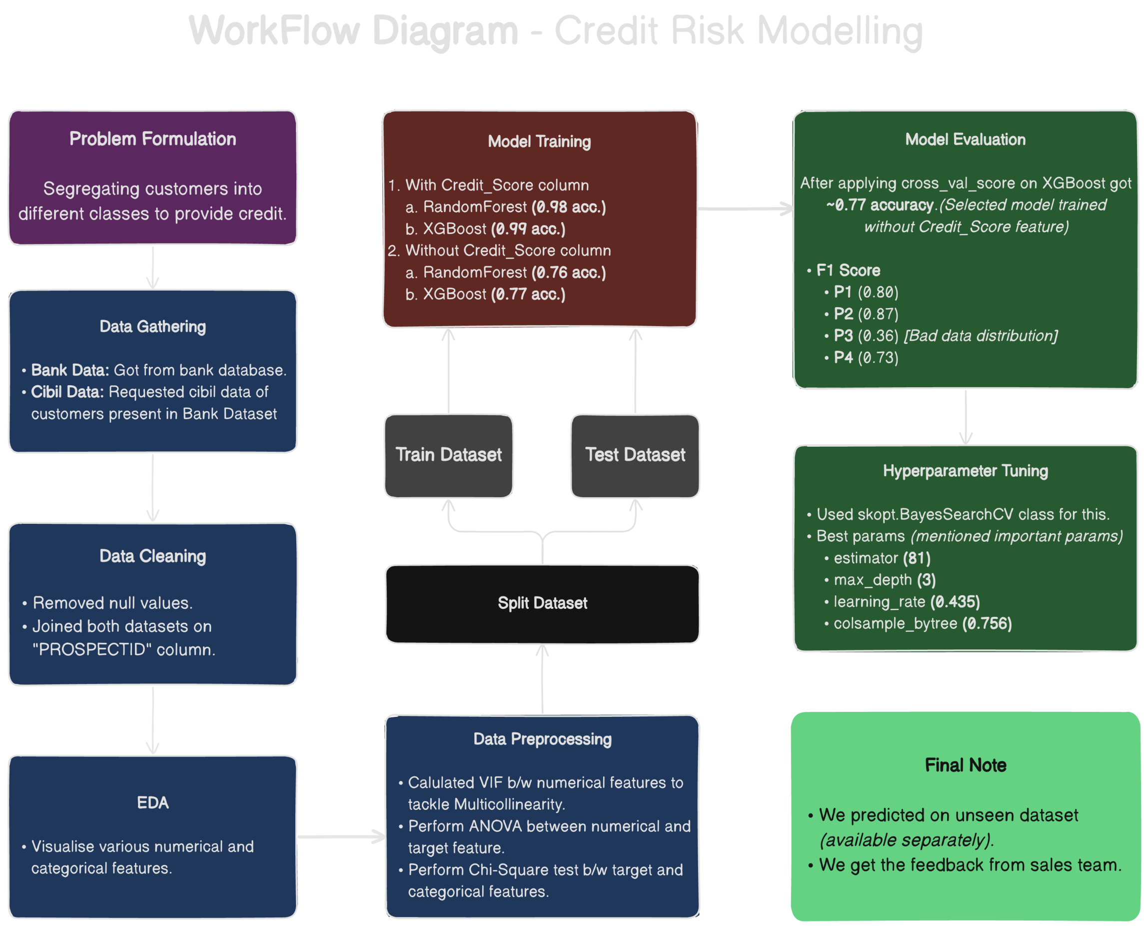
🏛 Credit Risk Modeling - Project

Problem Statement

Creating a machine learning model that can precisely segregate customers into class of giving credit based on past financial data and other pertinent borrower characteristics including income, credit score, and loan details is the aim. The likelihood that a borrower will fail on a loan should be estimated by the model, allowing it to determine the risk of lending to them. Such models can help financial institutions identify and measure their total risk exposure, set appropriate risk limits, and make informed investment decisions.

Project WorkFlow



Challenges Faced

One of the biggest challenges in credit risk modelling is the limited availability of relevant and reliable data. Credit risk models require historical data on loan performance, default rates, and economic indicators to accurately assess the likelihood of default.

Challenges include data availability, data quality, complex modelling, and regulatory compliance.

Example: One common challenge faced by financial institutions is obtaining accurate and reliable data for credit risk modelling purposes.

Detail Model Description

1. There are two datsets. We need to solve the challenges that are faced by the bank during credit lending.
2. First dataset is (i) Internal bank dataset and second dataset is (ii) Civil external dataset.
3. The target variable is Approved_Flag which contain 4 categories ['P2', 'P1', 'P3', 'P4'], segregating the customer into class of giving the credit. P1 being the category where the bank can easier give the credit to that customer whereas P4 being the category where it is not a good idea to give the credit to that customer, as it can increase the NPA accounts(Non-Performing assets) of the bank.
4. There are total 84 columns in two datasets. 26 columns in the first dataset and 62 columns in the second dataset.
5. `PROSPECTID` col is a common column in both the first and second datasets indicating unique customer ID.
6. To find association between numerical and numerical columns we will perform VIF test (Variance Inflation Factor). Reject columns whose p value is greater than a particular threshold.
7. For feature selection, we will perform Chi2-Square test and ANOVA test, since the target column is multi-class categorical column.
8. By checking the p_value of each column w.r.t target variable, we can decide if it's statistically significant or not.
9. Made two models. One without credit score and another with credit score.
10. It is observed that the accuracy of model without credit score feature has dramatically decreases.
11. Without credit score the accuracy is 77% and with credit score the accuracy is 99%.

Dataset Columns Description

Bank Dataset

Column	Description
`pct_tl_open_L6M`	Percent accounts opened in last 6 months
`pct_tl_closed_L6M`	percent accounts closed in last 6 months
`Tot_TL_closed_L12M`	Total accounts closed in last 12 months
`pct_tl_closed_L12M`	percent accounts closed in last 12 months
`Tot_Missed_Pmnt`	Total missed Payments
`CC_TL`	Count of Credit card accounts
`Home_TL`	Count of Housing loan accounts
`PL_TL`	Count of Personal loan accounts
`Secured_TL`	Count of secured accounts
`Unsecured_TL`	Count of unsecured accounts
`Other_TL`	Count of other accounts
`Age_Oldest_TL`	Age of oldest opened account
`Age_Newest_TL`	Age of newest opened account

Civil Dataset

Column	Description
`time_since_recent_payment`	Time Since recent Payment made
`max_recent_level_of_deliq`	Maximum recent level of delinquency

Column	Description
`num_deliq_6_12mts`	Number of times delinquent between last 6 and last 12 months
`num_times_60p_dpd`	Number of times 60+ dpd
`num_std_12mts`	Number of standard Payments in last 12 months
`num_sub`	Number of sub standard payments - not making full payments
`num_sub_6mts`	Number of sub standard payments in last 6 months
`num_sub_12mts`	Number of sub standard payments in last 12 months
`num_dbt`	Number of doubtful payments
`num_lss`	Number of doubtful payments in last 12 months
`recent_level_of_deliq`	Number of loss accounts in last 12 months
`CC_enq_L12m`	Credit card enquiries in last 6 months
`PL_enq_L12m`	Personal Loan enquiries in last 6 months
`time_since_recent_enq`	Personal Loan enquiries in last 12 months
`enq_L3m`	Enquiries in last 6 months
`last_prod_enq2`	Lates product enquired for
`first_prod_enq2`	First product enquired for
`MARITALSTATUS`	Marital Status
`EDUCATION`	Education level
`AGE`	Age

Column	Description
`GENDER`	Sex
`Time_With_Curr_Empr`	Time with current Employer
`CC_Flag`	Credit card Flag
`PL_Flag`	Personal Loan Flag
`pct_PL_enq_L6m_of_ever`	Percent enquiries PL in last 6 months to last 6 months
`pct_CC_enq_L6m_of_ever`	Percent enquiries CC in last 6 months to last 6 months
`HL_Flag`	Housing Loan Flag
`GL_Flag`	Gold Loan Flag
`Approved_Flag`	Priority levels

Important Notes From Both The Dataset

1. The shape of bank internal dataset of customer is (51336, 26).
2. The shape of civil dataset is (51336, 62)
3. The common column in both datset is `PROSPECTID` which is unique ID for each customer.
4. The value "-99999" in both the datasets are null values.
5. We will remove all the null values if data lost is less than 20% of the total dataset.
6. Total trade lines is total no of accounts of a customer.

EDA

Unique values in categorical columns

Column	Unique Values
MARITALSTATUS	Married, Single

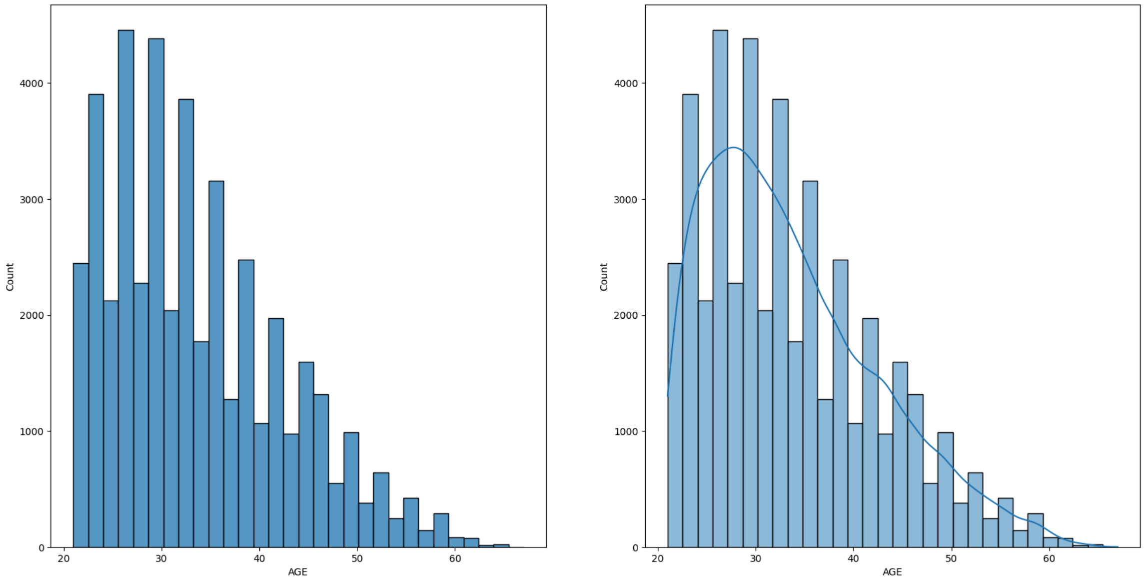
Column	Unique Values
EDUCATION	12TH, GRADUATE, SSC, POSTGRADUATE, UNDERGRADUATE, OTHERS, PROFESSIONAL
GENDER	M, F
last_prod_enq2	PL, ConsumerLoan, AL, CC, others, HL
first_prod_enq2	PL, ConsumerLoan, others, AL, HL, CC
Approved_Flag	P2, P1, P3, P4

After performing all the statistics tests (i.e. Chi-Square, VIF and ANOVA test), it is found that only 43 columns are important out of 82 columns.

```
[
'Age_Newest_TL', 'Age_Oldest_TL', 'Approved_Flag', 'CC_enq_L12m', 'CC_Flag',
'CC_TL', 'EDUCATION',
'first_prod_enq2', 'GENDER', 'GL_Flag', 'HL_Flag', 'Home_TL',
'last_prod_enq2', 'MARITALSTATUS',
'max_recent_level_of_deliq', 'NETMONTHLYINCOME', 'num_dbt_12mts', 'num_dbt',
'num_deliq_6_12mts',
'num_lss', 'num_std_12mts', 'num_sub_12mts', 'num_sub_6mts', 'num_sub',
'num_times_60p_dpd',
'pct_CC_enq_L6m_of_ever', 'pct_PL_enq_L6m_of_ever', 'pct_tl_closed_L12M',
'pct_tl_closed_L6M',
'pct_tl_open_L6M', 'PL_enq_L12m', 'PL_Flag', 'PL_TL', 'recent_level_of_deliq',
'Secured_TL',
'time_since_recent_enq', 'time_since_recent_payment', 'Time_With_Curr_Empr',
'Tot_Missed_Pmnt',
'Tot_TL_closed_L12M', 'Unsecured_TL', 'enq_L3m', 'Other_TL',
]
```

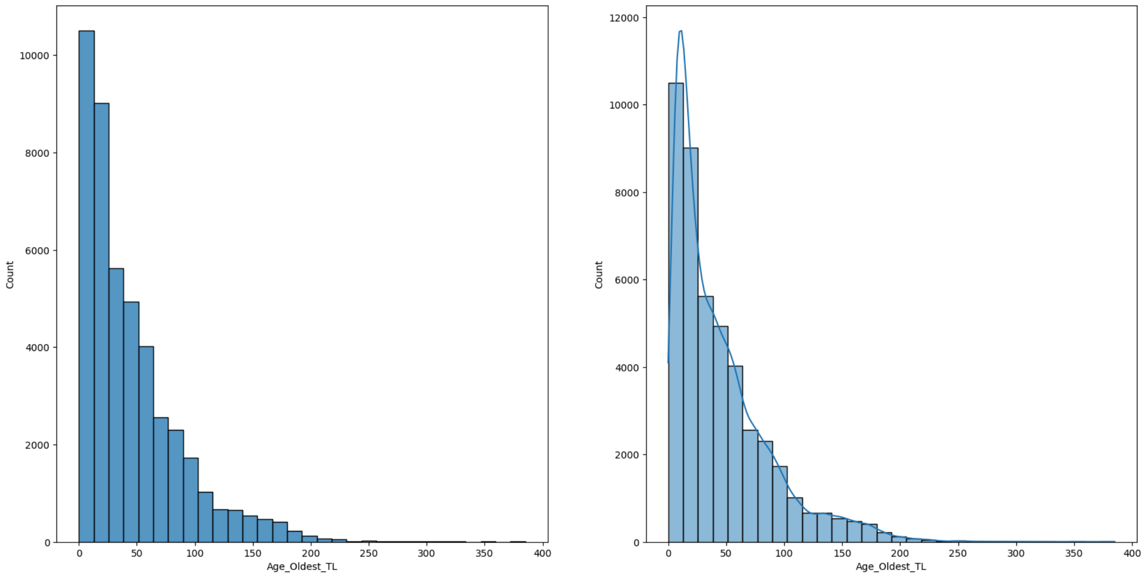
Data Visualization

Age Distribution Graph

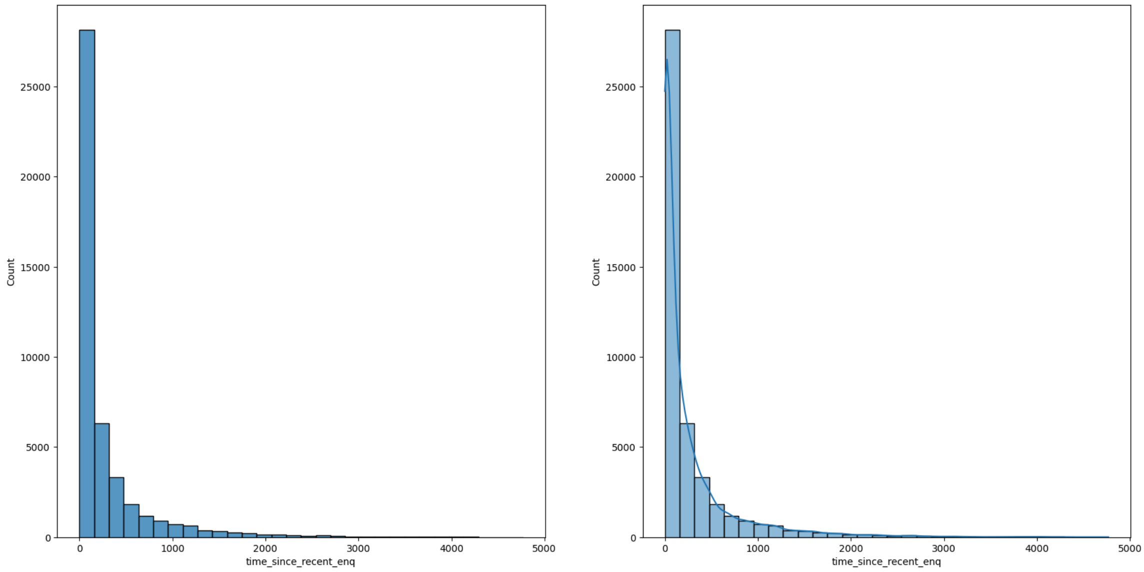


- The bank is majorly targeting people between 20 to 40.
- The data distribution of this dataset is majorly spread between the age group of 20 to 40.

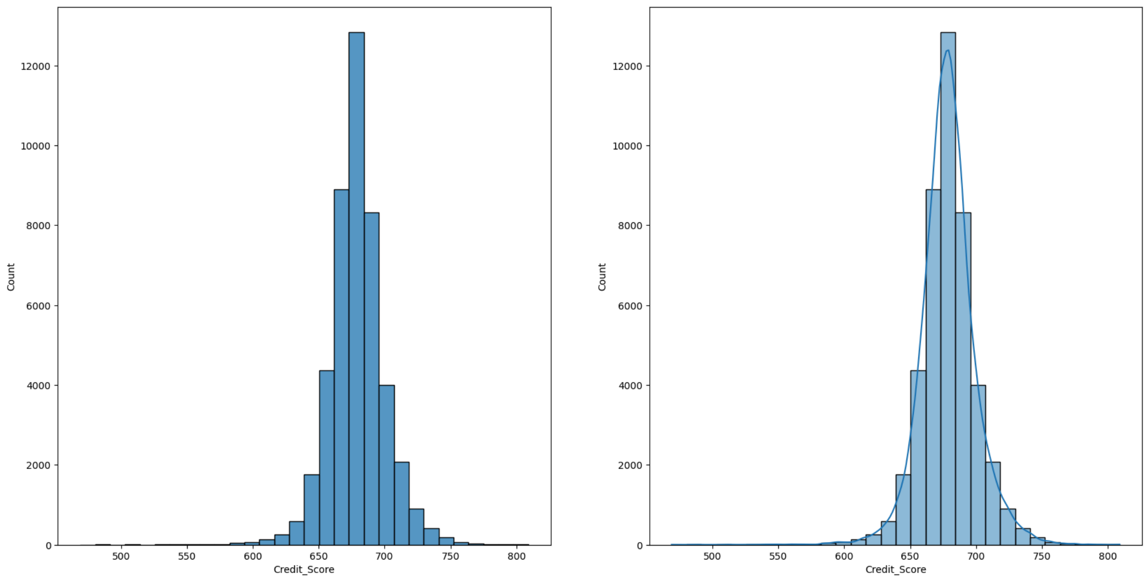
Age of Oldest loan/Trade Line account (In Months)



Time since recent enquiry (In Months)

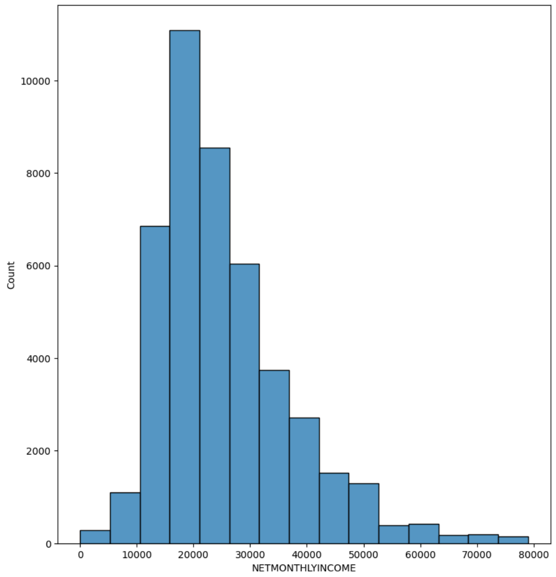


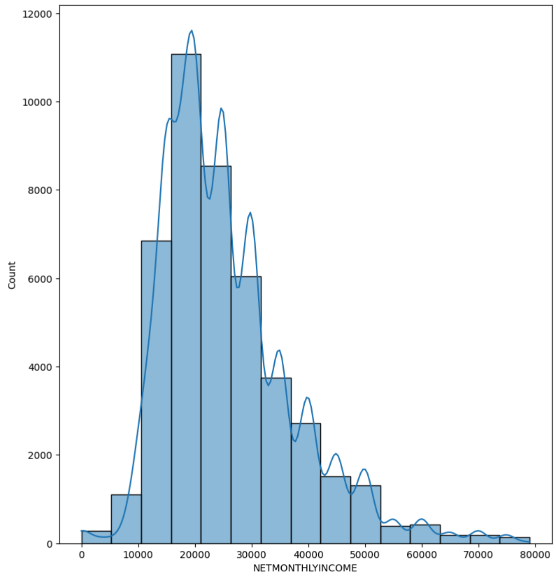
Credit Score Distribution



Most of the data distribution of credit score column is spread between 660 and 700 which fall under P2 category and that's why majorly category in target column is P2 category only.

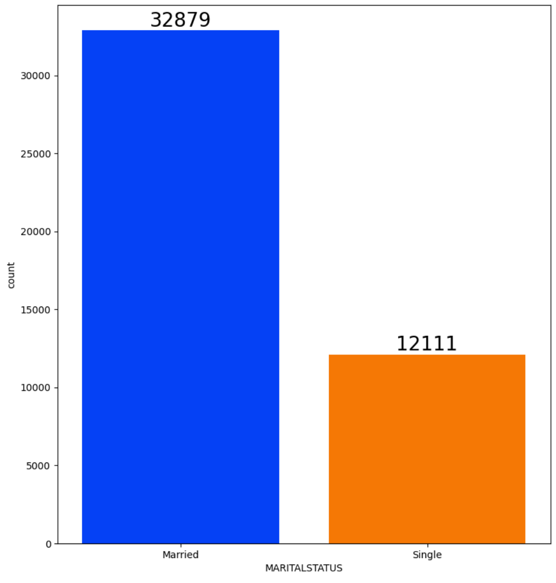
90 percentile Monthly Income Distribution

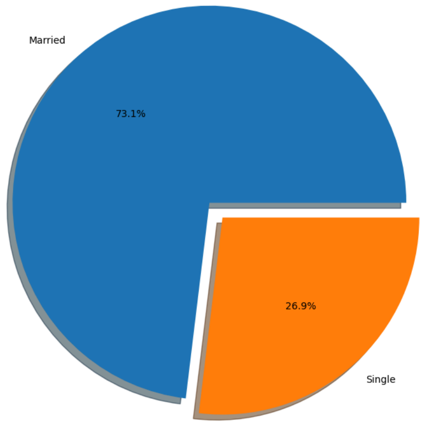




This column illustrate that the salary income majority of people falls between 20k to 35k. It can be observed that the bank is mainly targeting those people whose is under 50k per month

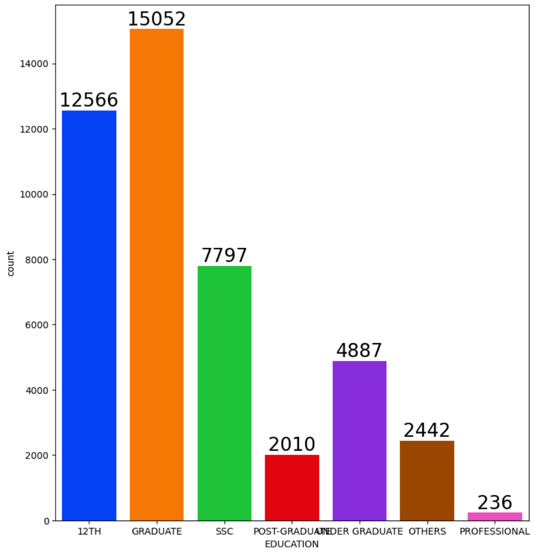
Marital Status Distribution Graph

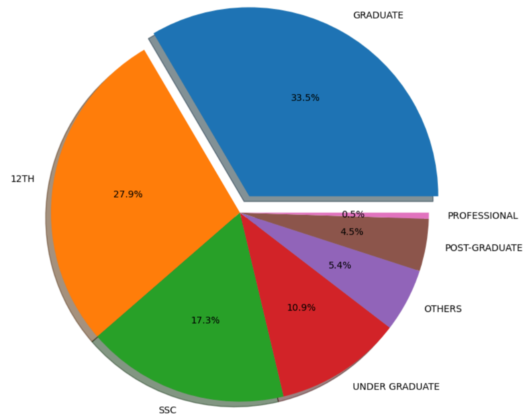




This column indicates that 73.1% of people who are applying for the loan are married.

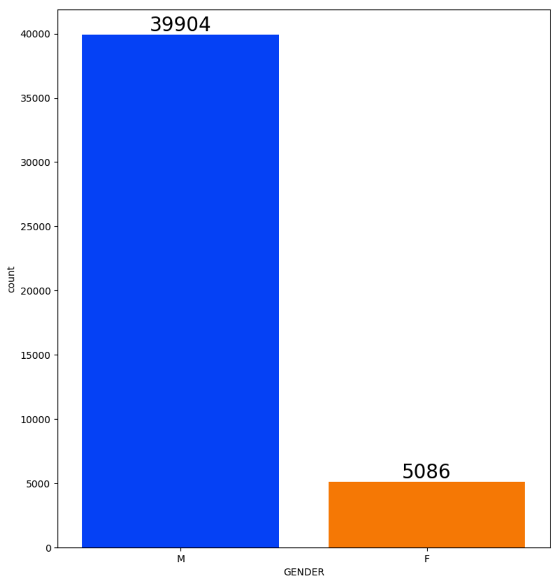
Education Distribution Graph

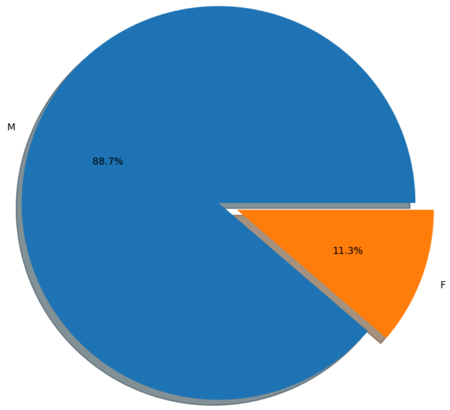




The Graduate and 12th pass population contribute significantly to the dataset

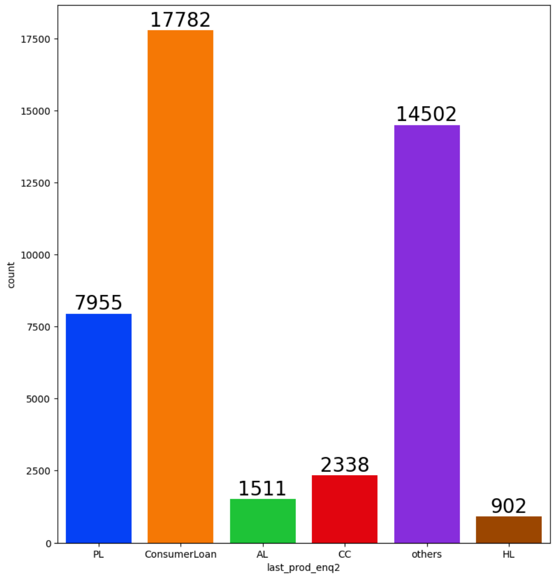
Gender Distribution Graph:

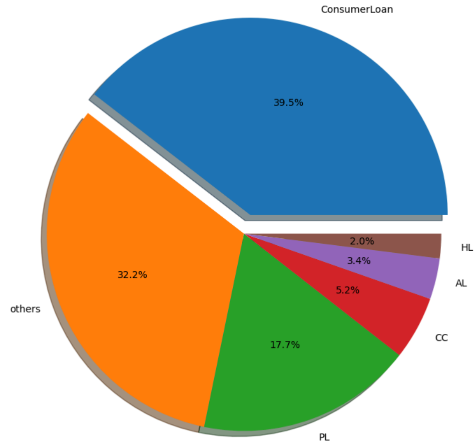




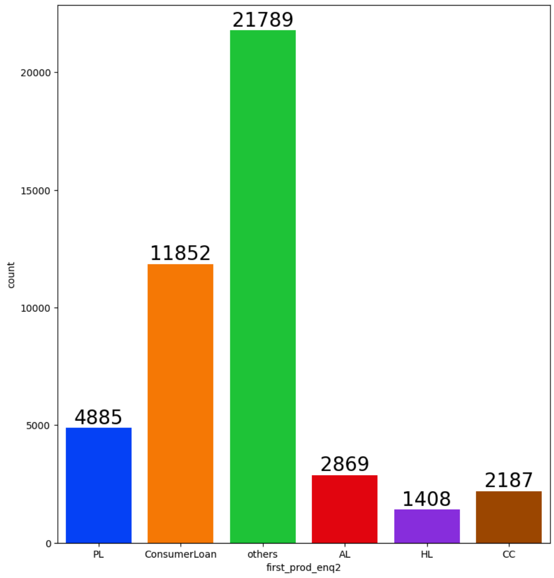
Gender wise, the dataset shows that 88.7% who are applying for loan are male or we can say that the bank is targeting male candidate more.

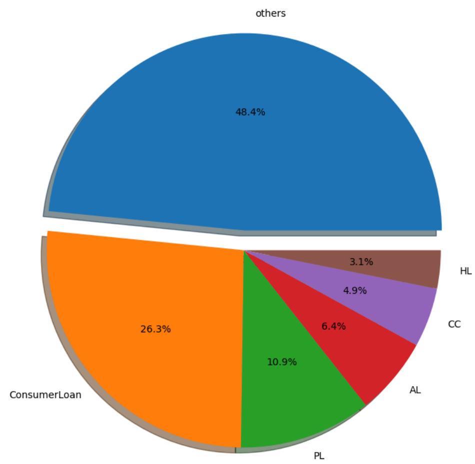
Last Product Enquiry Graph



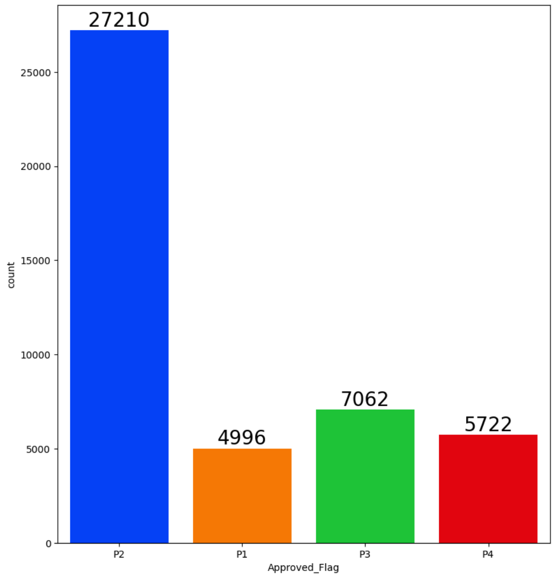


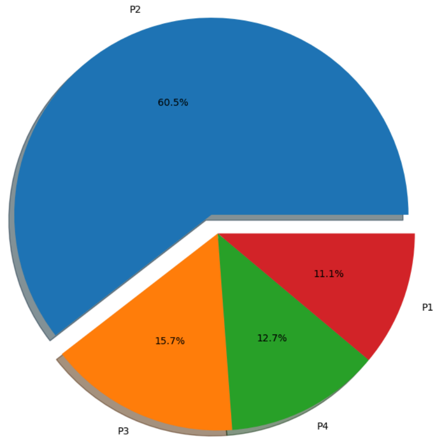
First Product Enquiry Graph





Distribution of Target Variable Categories



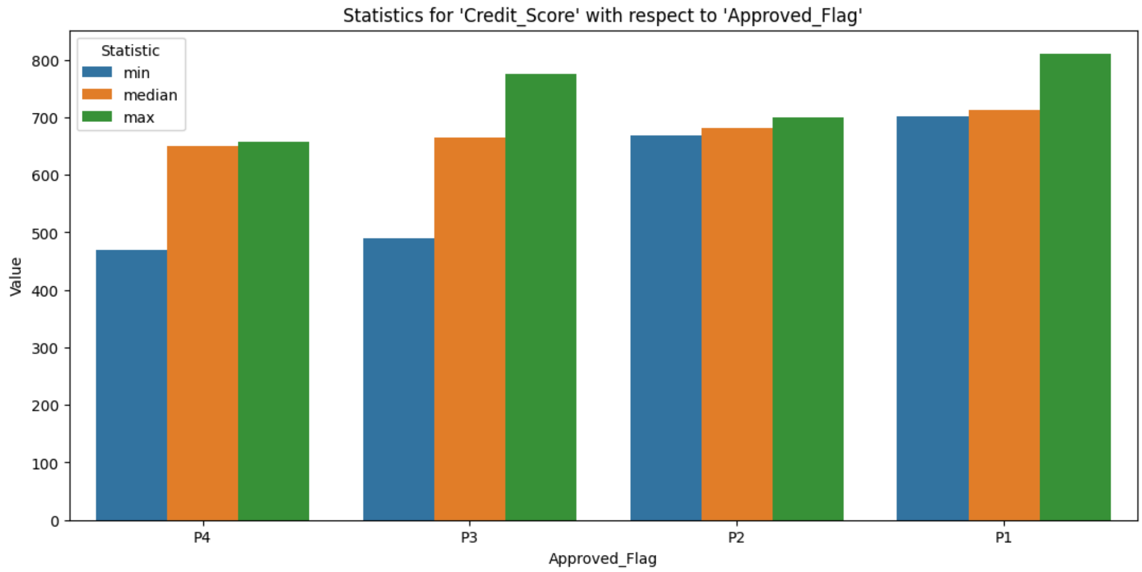


60% of the people in the dataset falls under P2 category for loan approval.

Data Visualization - Observations

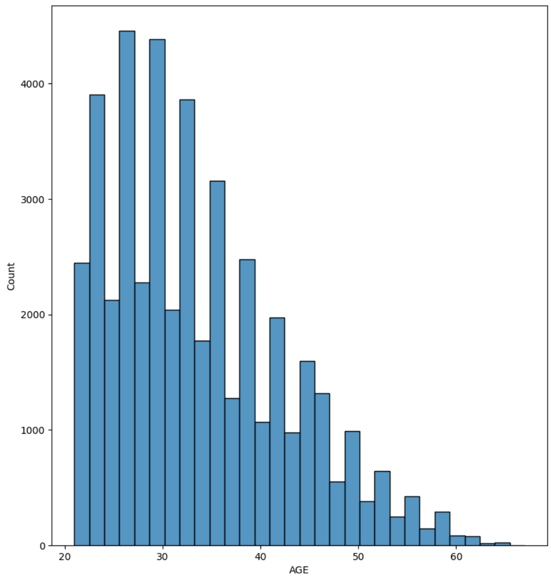
1. 73% people are married people in the dataset.
2. This datset contains have 88% of people as men who are likely to taken loan from the bank.
3. Graduate people have more likelihood of taking or applying for loans. Banks also have more likelihood of approving of loans to the graduate or educated people.
4. Previous loans taken by the people in this dataset is other loan or consumer loans (such as furniture loan, fridge loan etc).
5. Most of the people in the dataset flows under P2 category for loan approval.

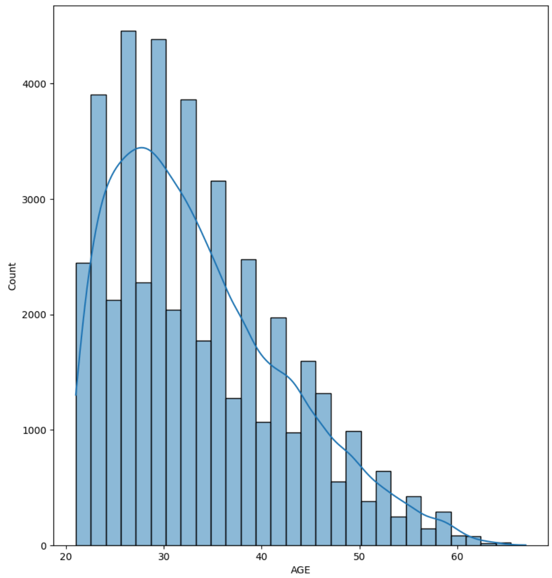
Minimum, Maximum and Median value of `Credit_Score` for each category



The min and max credit score is P3 category is 489 and 776 respectively. This range indicates that for P3 category creates a big ambiguity for the model to predict the output accurately. For P1 and P2 categories, it is easier for the model to predict as it range from (701, 809) and (689 and 700) respectively.

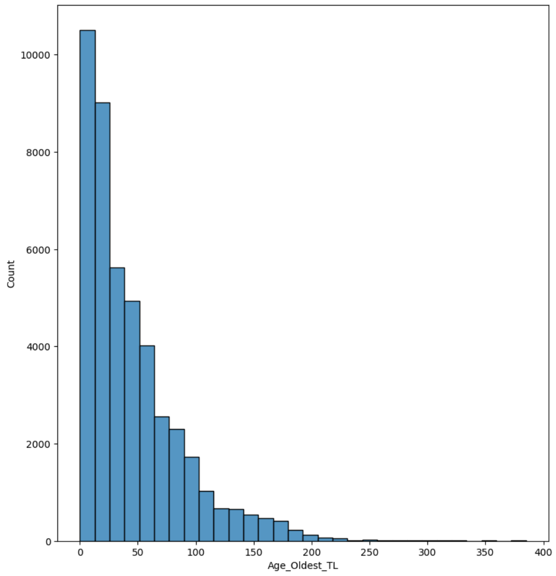
Minimum, Maximum and median value of `AGE` for each category

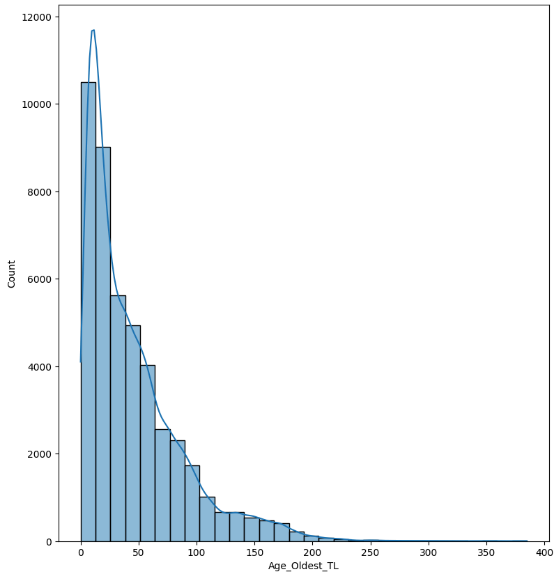




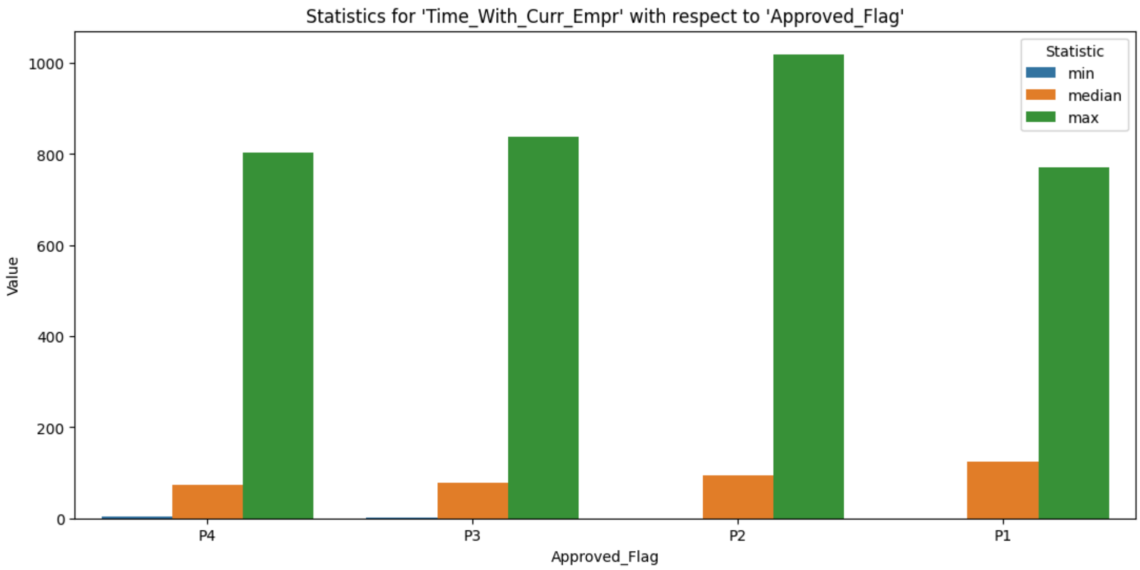
Min age for all the category is 21. Max age varies from 63 to 67. It can be observed that for P1 category the median age is higher as compare to other categories and as the category decrease median age also decreses.

Maximum and median value of Age Oldest Trade Line accounts for each category





Maximum and median value of Time with current enquiry accounts for each category



Observation of numerical and categorical cols w.r.t target variable i.e. (Approved_Flag)

1. P1 category range is (701-809)
2. P2 category range is(669-700)
3. P3 category range is (489-776)
4. P3 category of target variable are the most ambiguous category. This can be observed by looking at the credit score min and max value for P3 category which range from 489 to 776, whereas in case of P2 it's ranges from 669 to 701.

5. Due to the most ambiguous category i.e. P3, during the predict also, the accuracy of the model is significantly decreases due to the most ambiguous category.
6. The median age who are getting P1 category loan are bit older than other categories. For eg. median age for P1 category is 40 whereas for P2 category it is 33 and for P3 category it is 31. Therefore it can be assumed that as the age increases, loan approval becomes easier.

Model Training

We used ensemble techniques (bagging and boosting) to train the model. Mainly we used **RandomForestClassifier** and **XGBoostClassifier** for classification. However, it is observed that **XGBoost** classifier has better accuracy as compare to **RandomForest** classifier. With accuracy 99% **XGBoost** is the best ML algorithm for the dataset with credit score feature is included. However, when credit score feature is excluded, there is a significant drop in the accuracy(76%) because the P3 category is most ambiguous category, the accuracy of the model is significantly decreases.

Classification Report of `RandomForest`

Using `Credit_Score` feature

	precision	recall	f1-score	support
P1	0.94	1.00	0.97	1224
P2	1.00	1.00	1.00	6397
P3	1.00	0.95	0.98	1595
P4	1.00	1.00	1.00	1309
accuracy			0.99	10525
macro avg	0.99	0.99	0.99	10525
weighted avg	0.99	0.99	0.99	10525

Without `Credit_Score` feature

	precision	recall	f1-score	support
P1	0.82	0.70	0.75	1224
P2	0.79	0.93	0.86	6397
P3	0.45	0.21	0.28	1595
P4	0.75	0.70	0.72	1309
accuracy			0.77	10525
macro avg	0.70	0.63	0.65	10525
weighted avg	0.74	0.77	0.74	10525

Classification Report of `XGBoost`

Class	P1	P2	P3	P4
Precision	0.813	0.826	0.434	0.772
Recall	0.788	0.912	0.305	0.698
F1 Score	0.800	0.866	0.358	0.733

Hyperparameter Tuning

Using `skopt.BayesSearchCV`, we got the best parameter for our XGBoost model. The parameters are:

```
{
"alpha": 10,
"colsample_bytree": 0.9,
"learning_rate": 1.0,
"max_depth": 3,
"n_estimators": 100,
"num_classes": 4,
"objective": "multi:softmax",
}
```

With these parameters, accuracy increases by 1% when the model is trained without `Credit_Score` feature.