Topic: Breaking Barriers - Micro Mortgage Analytics

Introduction

Data-driven decision making carries utmost importance in today's age. It not only helps businesses recognize the areas of improvement but also solves critical problems. Business analytics helps organisations find trends and underlying patterns in data from various aspects, build targeted marketing action plans, predict the problematic situations that can arise shortly to take preventive measures, etc. Shubham Housing Development Finance Company caters to a different segment of customers who do not fall under the conventional loan applicants category. Their loan approval process involves intense field work and high expertise in analysing the customer's creditworthiness.

Significance of Modelling in Business

Shubham Housing Development Finance Company is the pioneer in serving loans to the informal sector, which made their business expand rapidly in 2 years, having 40 branches across the country. However, this has posed its own set of challenges. Since their loan process involves a lot of fieldwork and customer analysis, there is a lot of time and cost involved to process each applicant. Also, there is a challenge related to the standardisation of the approval process, because the whole interview and assessment are based on individual credit officers' perceptions. This has been the major driving factor for Shubham Housing Development Finance Company in acknowledging the importance of data-driven decision-making, and they believe that this will help them to grow their business and stay ahead of the competition.

Why is the “Branch Level Credit Check” important in Modeling?

Out of all the steps involved in the approval process, “Branch Level Credit Check” is the preliminary step that finalises the involvement of credit officers and the rest of the loan process. Since the company invests a lot of time and effort in analysing every loan request through fieldwork and assessment, it is necessary to examine the applications and evaluate the possibility of rejecting them well before investing a lot of resources. This will help the company

to divert the resources to other applicants who have a high chance of loan approval, thereby reducing the lead times in the approval process as well as various costs associated with it.

How Modeling is Helping the Company?

As discussed earlier, modelling not only helps to reduce the costs, but it also helps to standardise their decision-making process across the country. Through modelling, they can reach out to potential credit-worthy customers and at the same time reduce the possibility of giving loans to clients who can be risky in the future.

Data analysis was carried out using the information provided in the exhibits. We can understand the nature of the data that is being collected. At the same time, one can analyse the correlation between various parameters, which can impact the decision-making process. The derived data from “Exhibit-6” turned out to be most valuable for the decision-making process upon our analysis.

Exploratory data analysis has been carried out on the given information and below are a couple of observations:

- Most no.of the applicants are female, with U10 education background.
- Most female applicants have an education background of under 10th and applied without necessary documents.
- Correlation among the various attributes given in the exhibits is very minimal.
- Applicants having high LTV are more prone to rejections than applicants having low LTV. There aren’t any applicants with LTV higher than 90 who have got the loan sanctioned, serving as a threshold.
- Similarly applicants having four more than 50 have very high chances of rejection.
- In Tier 2 cities the average LTV is lower than the total average. They also have the highest acceptance rate.
- Applicants having an EMI amount contributing to 25-50% of Total household income have more chances to get the loan.
- Most of the applicants who fall between age groups 30-38 have good chances of approval.

Segmentation

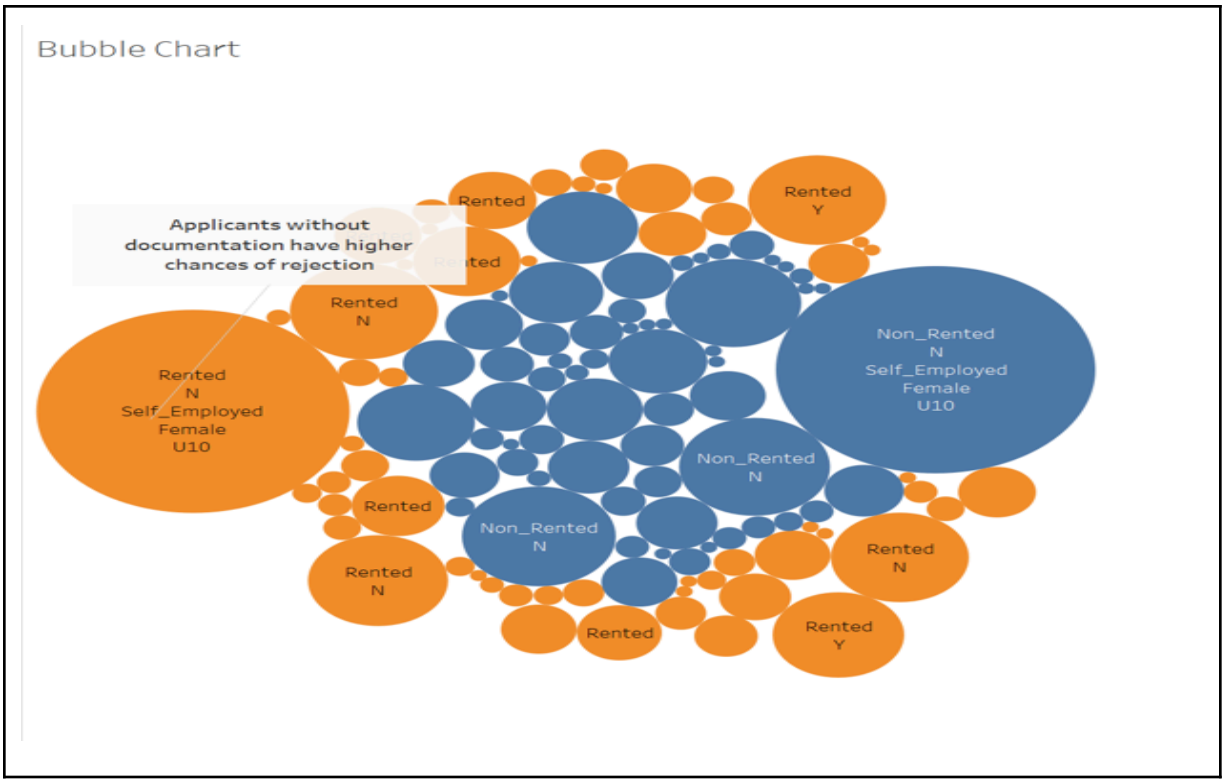
Accommoda..	Employment T..	Education Class	Doc Proof Inc / Employer Type N Busi..	N Corp..	N Govt	N Ind/..	Y Busi..	Y Corp..	Y Govt	Y Ind/..
Non_Rented	Salaried	GRADUATE+		37	11	17		27	5	21
		U10		42	27	89		35	47	52
		UNDERGRADUA..		24	8	21		19	20	23
	Self_Employed	GRADUATE+	74	1		1	6	2		
		U10	375	2		10	22	2	1	2
		UNDERGRADUA..	86	2		6	8	2		
Rented	Salaried	GRADUATE+		46	4	30		71	13	25
		U10		36	11	80		20	18	66
		UNDERGRADUA..		18	7	23		22	6	22
	Self_Employed	GRADUATE+	81	1	1	2	11	1		
		U10	358	1		13	27			2
		UNDERGRADUA..	71	1		1	5			2

Analysis is carried out on the applicants based on different aspects like Employment type, Education, current housing, etc.

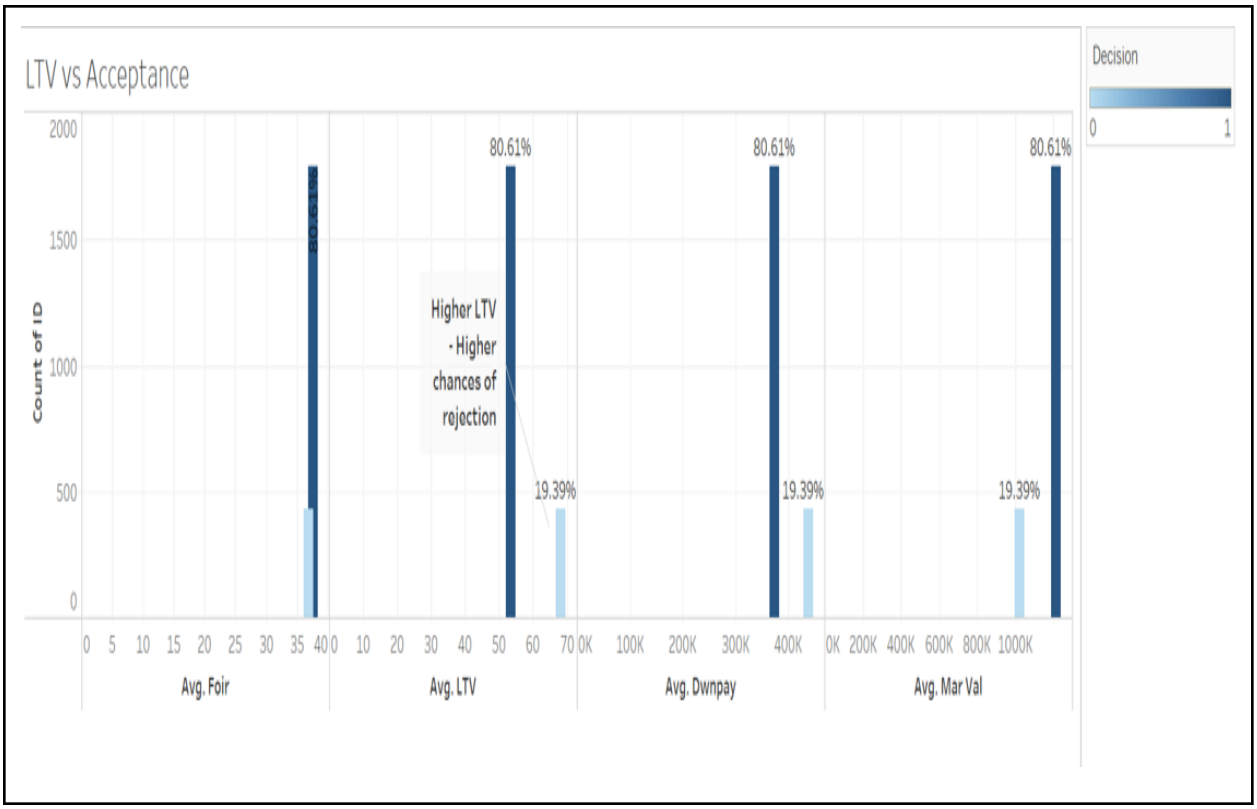
Overall Heatmap

Accommoda..	Employment T..	Gender	Employer T..	Doc Proof Inc / Education Class N GRADUATE+	N U10	N UNDERGRADUATE	Y GRADUATE+	Y U10	Y UNDERGRADUATE
Non_Rented	Salaried	Female	Corporate	7	15	2	2	10	2
			Govt	4	4	2	3	9	1
			Ind/SmallBus	3	11	5	4	8	3
		Male	Corporate	1			1		
			Govt		1				
			Ind/SmallBus	1	2			1	
	Self_Employed	Female	Business	22	97	29		4	2
		Male	Business	2	7	2		1	1
Rented	Salaried	Female	Corporate	5	8	2	7	4	3
			Govt			3	2	3	1
			Ind/SmallBus	1	6	1	4	6	3
		Male	Corporate	2			1		1
			Ind/SmallBus					1	
	Self_Employed	Female	Business	14	61	11		4	1
		Male	Business	2	4	1			

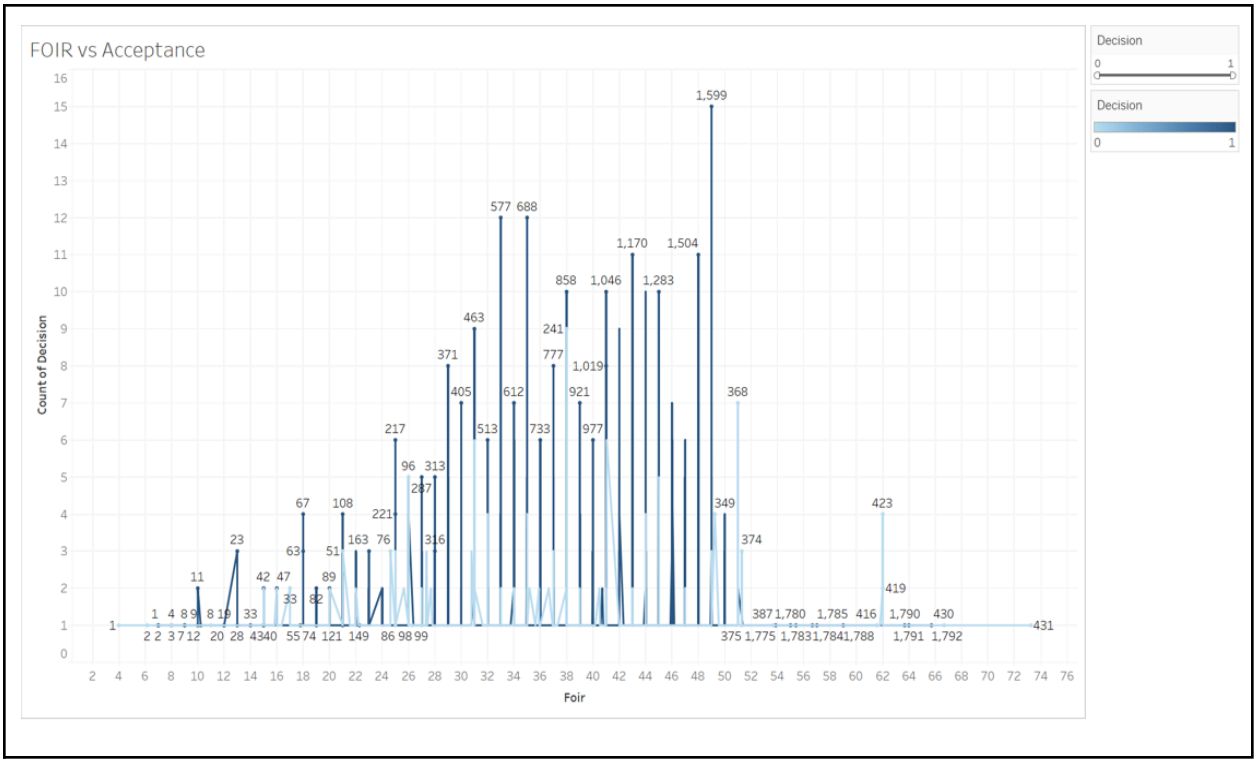
Category of applicants who got rejected most is highlighted with the darker colour.



Rejection rate is higher among applicants who don't contain basic documentation.

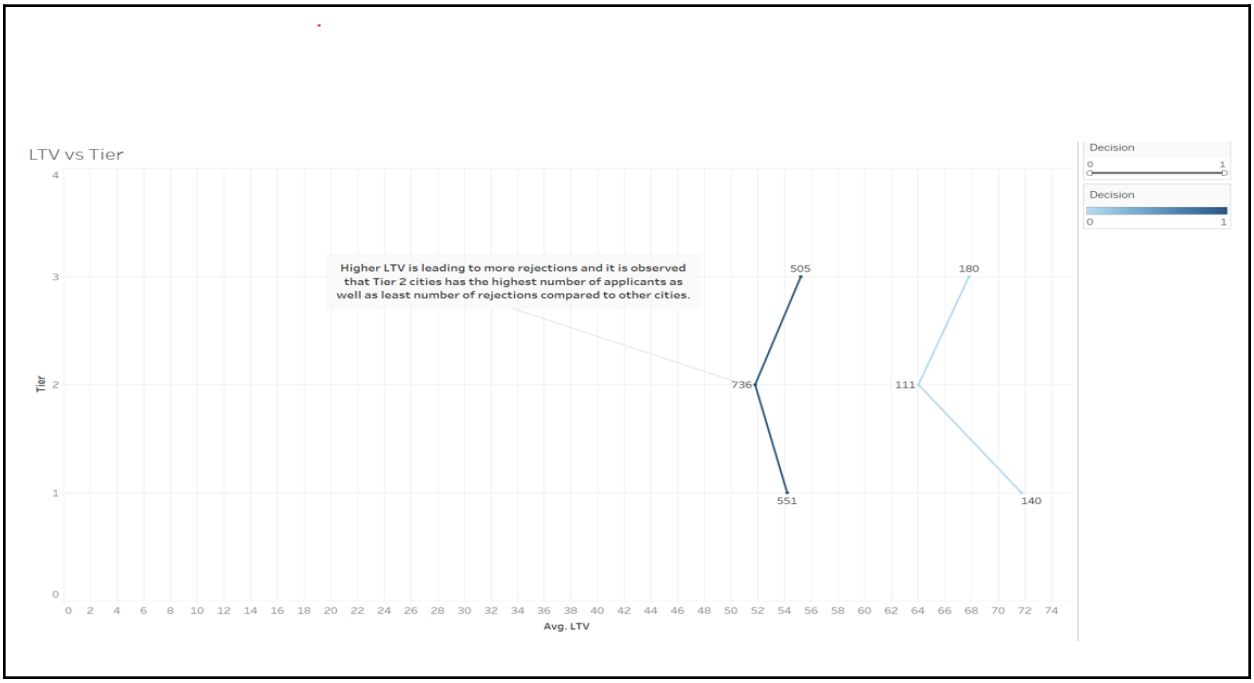


Higher LTV which is a derived factor leads to higher chances of rejection.

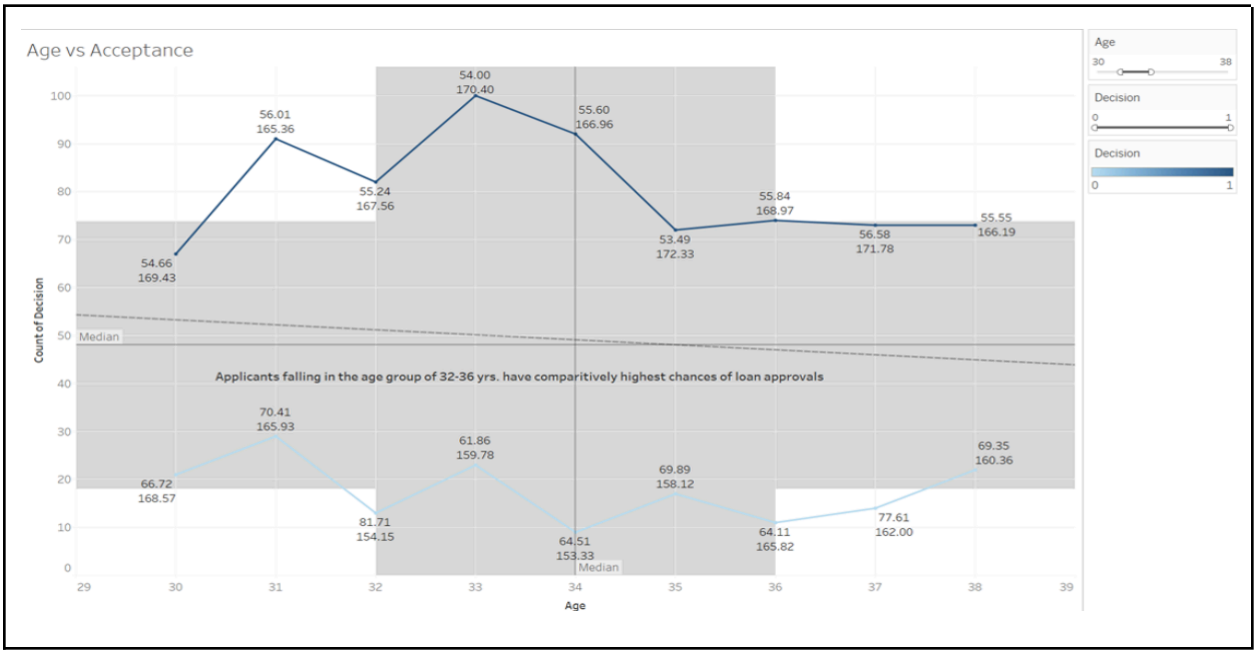


Decision-making can also be carried out based on derived attribute Foir. There are very few applications which got accepted whose Foir is higher than 50.

Note: The values shown above are cumulative/Running sum calculations.



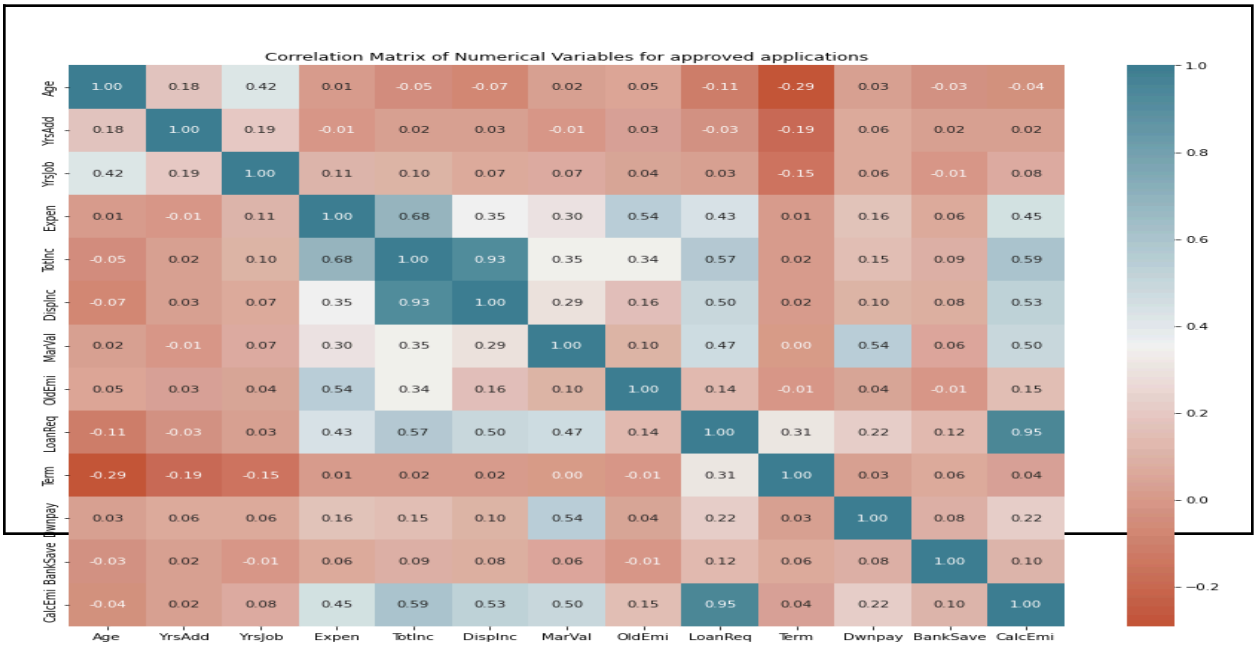
Most of the applications received are from Tier 2, and the rejection rate is lower in Tier 2 cities compared to others. Their average LTV is also lowest among Tier-1 and Tier-3 cities.

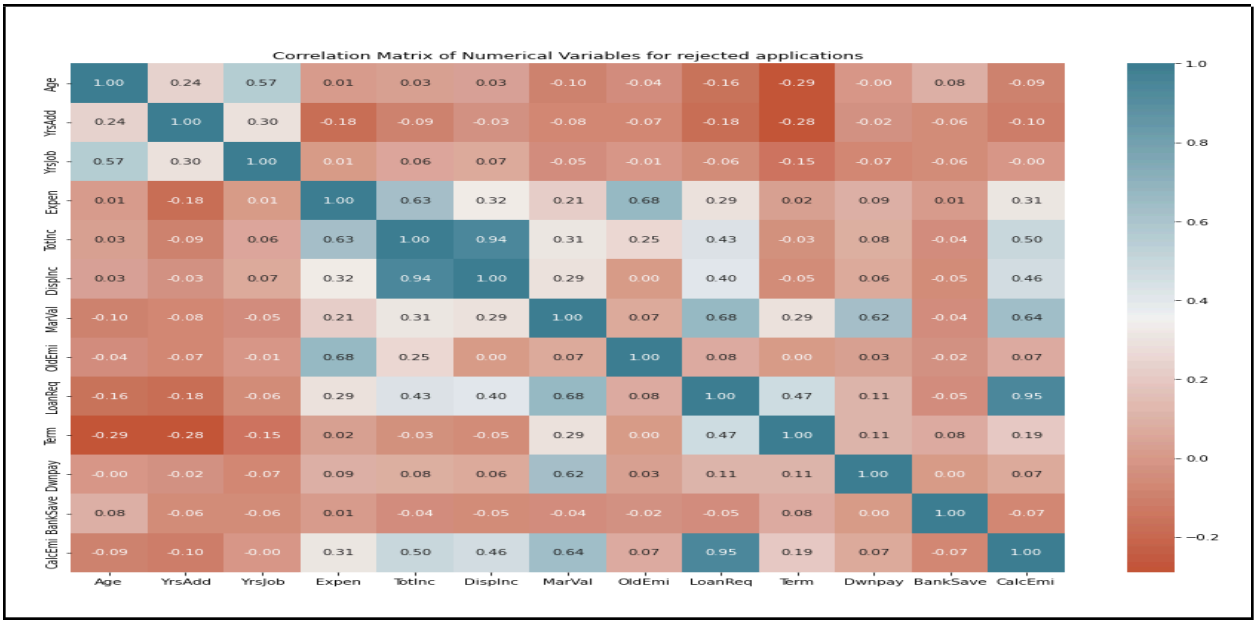


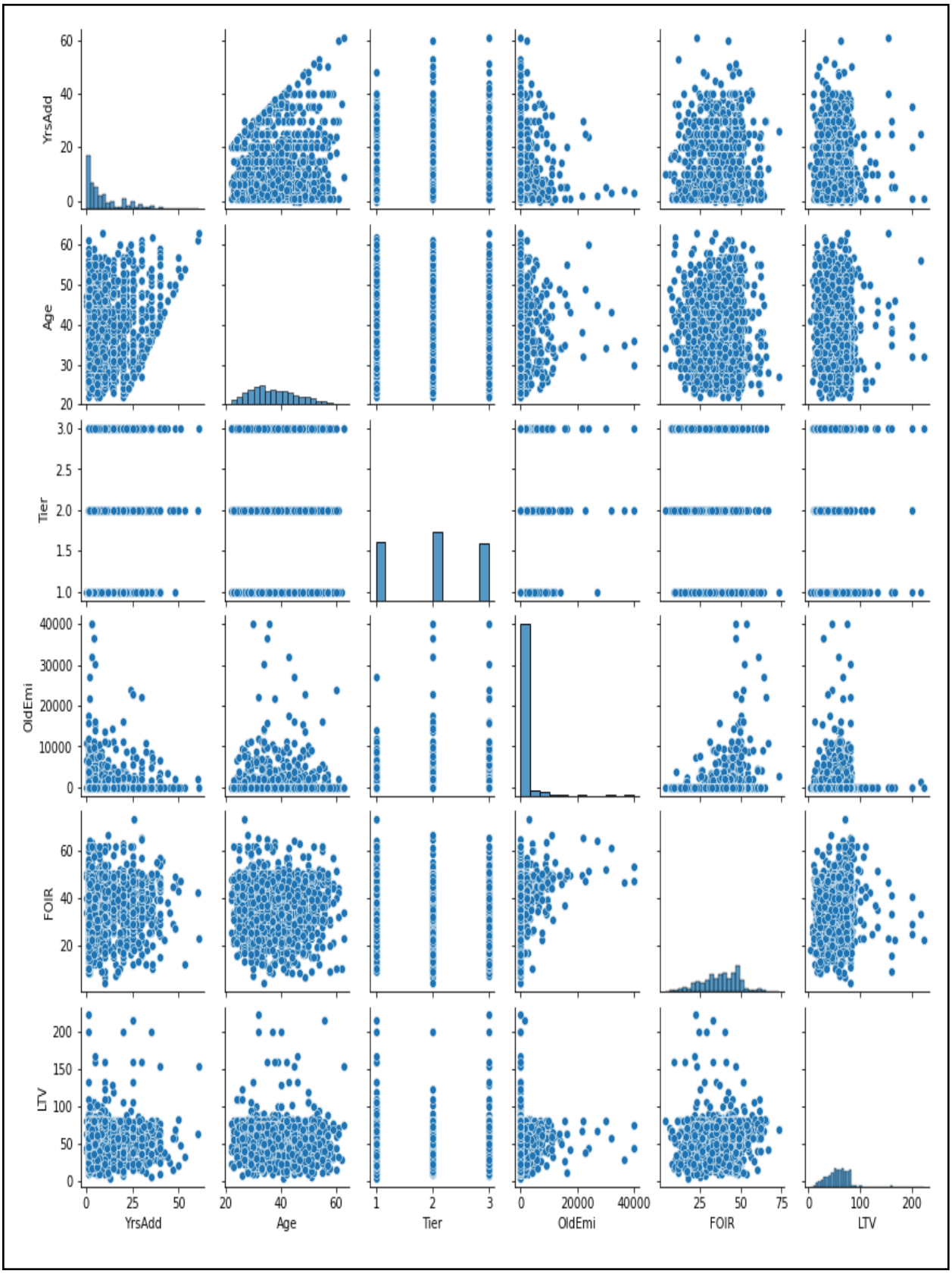
Upper quartile and lower quartile bands of the age group showing customers falling in 32-36 years have got the most approvals. (Their average LTV also falls in acceptable levels, those whose average LTV is high can be seen as rejected in the lower quartile region)

Based on the above observations, the company came up with a model using FOIR, Tier, Address at the present location, Age and old emi as some of the factors for decision-making.

Correlation plots between Numerical variables for accepted and rejected applications :




Correlation Matrix of Numerical Variables for rejected applications
Age 1.00 0.24 0.57 0.01 0.03 0.03 -0.10 -0.04 -0.16 -0.29 -0.00 0.08 -0.09
YrsAdd 0.24 1.00 0.30 -0.18 -0.09 -0.03 -0.08 -0.07 -0.18 -0.28 -0.02 -0.06 -0.10
YrsJob 0.57 0.30 1.00 0.01 0.06 0.07 -0.05 -0.01 -0.06 -0.15 -0.07 -0.06 -0.00
Expen 0.01 -0.18 0.01 1.00 0.63 0.32 0.21 0.68 0.29 0.02 0.09 0.01 0.31
TotInc 0.03 -0.09 0.06 0.63 1.00 0.94 0.31 0.25 0.43 -0.03 0.08 -0.04 0.50
DispInc 0.03 -0.03 0.07 0.32 0.94 1.00 0.29 0.00 0.40 -0.05 0.06 -0.05 0.46
MarVal -0.10 -0.08 -0.05 0.21 0.31 0.29 1.00 0.07 0.68 0.29 0.62 -0.04 0.64
OldEmi -0.04 -0.07 -0.01 0.68 0.25 0.00 0.07 1.00 0.08 0.00 0.03 -0.02 0.07
LoanReq -0.16 -0.18 -0.06 0.29 0.43 0.40 0.68 0.08 1.00 0.47 0.11 -0.05 0.95
Term -0.29 -0.28 -0.15 0.02 -0.03 -0.05 0.29 0.00 0.47 1.00 0.11 0.08 0.19
Dwnpay -0.00 -0.02 -0.07 0.09 0.08 0.06 0.62 0.03 0.11 0.11 1.00 0.00 0.07
BankSave 0.08 -0.06 -0.06 0.01 -0.04 -0.05 -0.04 -0.02 -0.05 0.08 0.00 1.00 -0.07
CalcEmi -0.09 -0.10 -0.00 0.31 0.50 0.46 0.64 0.07 0.95 0.19 0.07 -0.07 1.00
Age YrsAdd YrsJob Expen TotInc DispInc MarVal OldEmi LoanReq Term Dwnpay BankSave CalcEmi
1.0 0.8 0.6 0.4 0.2 0.0 −0.2


YrsAdd
Age
Tier
OldEmi
FOIR
LTV
0
25
50
20
40
60
1
2
3
0
20000
40000
75
100
200

Modelling Results:

We have created various linear and non-linear classification models and compared them on various performance metrics.

Validation Data					
	CART - Entropy	**CART - Gini**	**Logistic Regression**	**Random Forest Classifier**	**XGBoost Classifier**
Accuracy	0.74	0.76	0.83	0.84	0.83
F1 score	0.84	0.85	0.91	0.91	0.9
Precision	0.86	0.87	0.83	0.87	0.87
Recall	0.82	0.84	1	0.96	0.93
Specificity	0.38	0.4	0.09	0.33	0.38

Test Data - 100 Samples					
	CART - Entropy	**CART - Gini**	**Logistic Regression**	**Random Forest Classifier**	**XGBoost Classifier**
Accuracy	0.76	0.78	0.84	0.86	0.87
F1 score	0.85	0.86	0.91	0.92	0.92
Precision	0.85	0.88	0.84	0.89	0.91
Recall	0.85	0.84	0.99	0.95	0.94
Specificity	0.33	0.5	0.17	0.44	0.56

Appropriate Evaluation Metrics:

Striking the right balance between appropriate evaluation metrics is crucial as different evaluation metrics serve different purposes. The choice of metrics depends on the specific priorities and considerations of the business or organisation involved. Here's a brief overview of each metric and its relevance in this context:

1. **Accuracy**:
   - **Importance**: Accuracy is a common metric that represents the overall correctness of the model's predictions.
   - **Consideration**: While accuracy is an essential metric, it might not be sufficient on its own, especially in imbalanced datasets where one class (e.g., approved loans) dominates the other (e.g., rejected loans).

2. **Precision**:
   - **Importance**: Precision is the ratio of correctly predicted positive observations to the total predicted positives. In the context of loan approval, precision is crucial because it measures how many of the approved loan predictions are correct.
   - **Consideration**: A higher precision means fewer false positives, i.e., fewer instances where a loan is approved incorrectly.

3. **Recall (Sensitivity)**:
   - **Importance**: Recall, or sensitivity, is the ratio of correctly predicted positive observations to all observations in the actual positive class. In the context of loan approval, recall is important because it measures the ability of the model to capture all the approved loans.
   - **Consideration**: A higher recall means fewer false negatives, i.e., fewer instances where a loan that should have been approved is rejected.

4. **Specificity**:
   - **Importance**: Specificity is the ratio of correctly predicted negative observations to the total actual negatives. In the context of loan approval, specificity is important because it measures the ability of the model to correctly identify rejected loans.

- **Consideration**: A higher specificity means fewer false positives in the negative class, i.e., fewer instances where a loan that should have been rejected is approved.

5. **F1 Score**:
   - **Importance**: The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is particularly useful in imbalanced datasets.
   - **Consideration**: F1 score is beneficial when there is a trade-off between precision and recall, helping to find a balance that suits the specific goals of the loan approval system.

In the context of a loan application approval system, precision and recall are often considered more important than accuracy because misclassifying a loan application can have significant financial consequences. The choice between precision and recall depends on the business's priorities and the acceptable level of risk. If the cost of false positives (approving a risky loan) is high, emphasis might be placed on precision. If the cost of false negatives (rejecting a good loan) is high, emphasis might be placed on recall. Specificity is also crucial in this context as it directly relates to the correct identification of rejected loans. The F1 score provides a balanced metric that considers both precision and recall simultaneously.

Observation & Conclusion:

The following are the key observations that we observed from our modelling activity:

1. The CART model with the Gini impurity criterion exhibited superior evaluation metrics compared to the CART model utilising entropy.
2. We observed the Logistic Regression model demonstrates commendable recall scores and has relatively outperformed the other models on recall.
3. However, when comparing overall evaluation metrics, the ensemble models like Random Forest and XGBoost performed better than the linear ones.
4. And we can consider the champion model as XGBoost. As, it excels across all the other metrics, indicating its superior performance, making it the preferred choice for a decision model.

References

1. Rudravaram, J. (2013). Breaking Barriers: Micro-Mortgage Analytics. Harvard Business Publishing Education. https://hbsp.harvard.edu/product/IMB445-PDF-ENG
2. Course Material: DSM 404 - Business Analytics, Prof. Adiyta Maheshwari, IIM Indore.
3. The following models are used and referred to,
    a. **CART (Decision Tree):**
        i. Library: scikit-learn (Python)
        ii. Official Link: scikit-learn Decision Trees
    b. **Logistic Regression:**
        i. Library: scikit-learn (Python)
        ii. Official Link: scikit-learn Logistic Regression
    c. **Random Forest:**
        i. Library: scikit-learn (Python)
        ii. Official Link: scikit-learn Random Forest
    d. **XGBoost:**
        i. Library: XGBoost (Python)
        ii. Official Link: XGBoost Python Package