Applied Machine Learning

BUAN 6341.002

Group-03

Vamsi Krishna Kanderi Murali

Bhupesh Kumar Srivastava

Mayur Kumar Tikmani

Rajesh Bhattacharjee

Sonal Seth

1. Introduction, Business Problem, Motivation and Setting:

Commercial banks receive a lot of applications for credit cards. Many of them get rejected for many reasons, like high loan balances, low-income levels, or too many inquiries on an individual's credit report, for example. Manually analyzing these applications is mundane, error-prone, and time-consuming (and time is money!). Fortunately, this task can be automated with machine learning, and pretty much every commercial bank does so nowadays. In this project, we will build an automatic credit card approval predictor using machine learning techniques, just like real banks do. Predicting whether a credit card application will be approved or rejected based on values of feature variables is a supervised machine learning classification task. We plan to use the logistic regression model for this problem.

2. Data Description:

The dataset contains the following information: the personal and financial information of a person who has applied for credit card approval. Each row represents an applicant and whether their application was approved or not.

No. of records: 25128

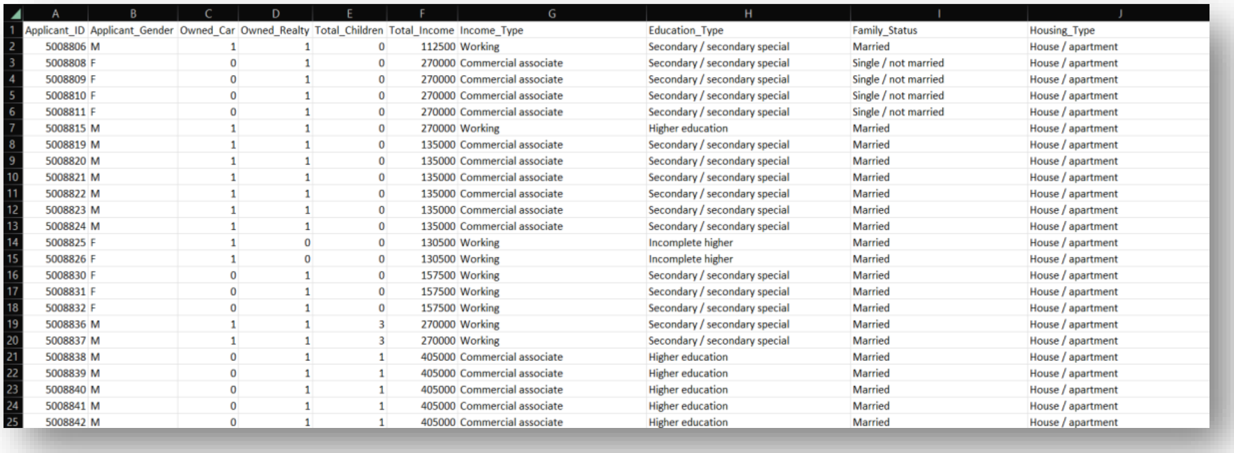
No. 0f Unique variables: 21

Classes: 121 – Approved, 25007 - Not Approved

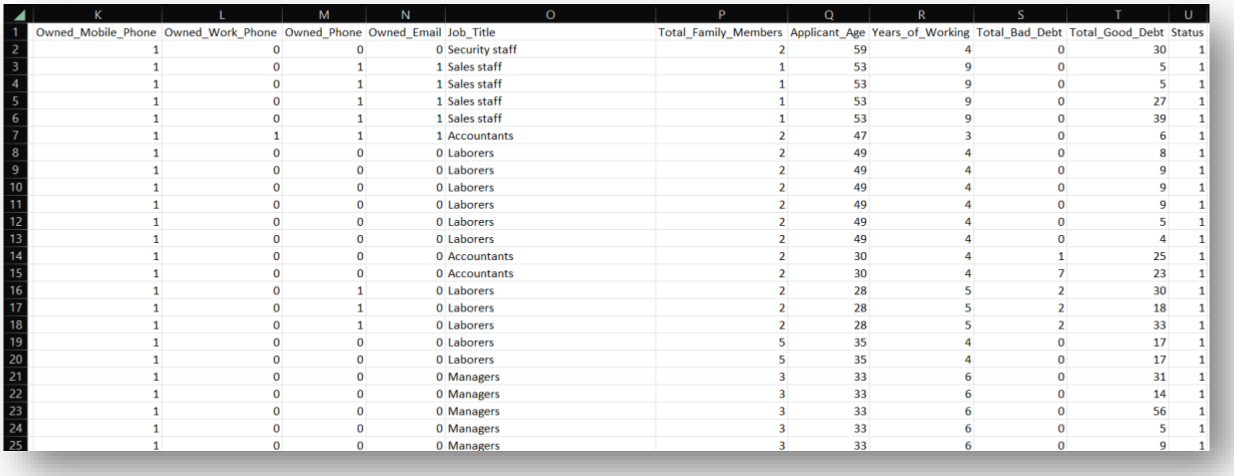
Total Null/Missing values: 0

Columns Type: Numerical & Categorical

Columns with highest missing values: N/A



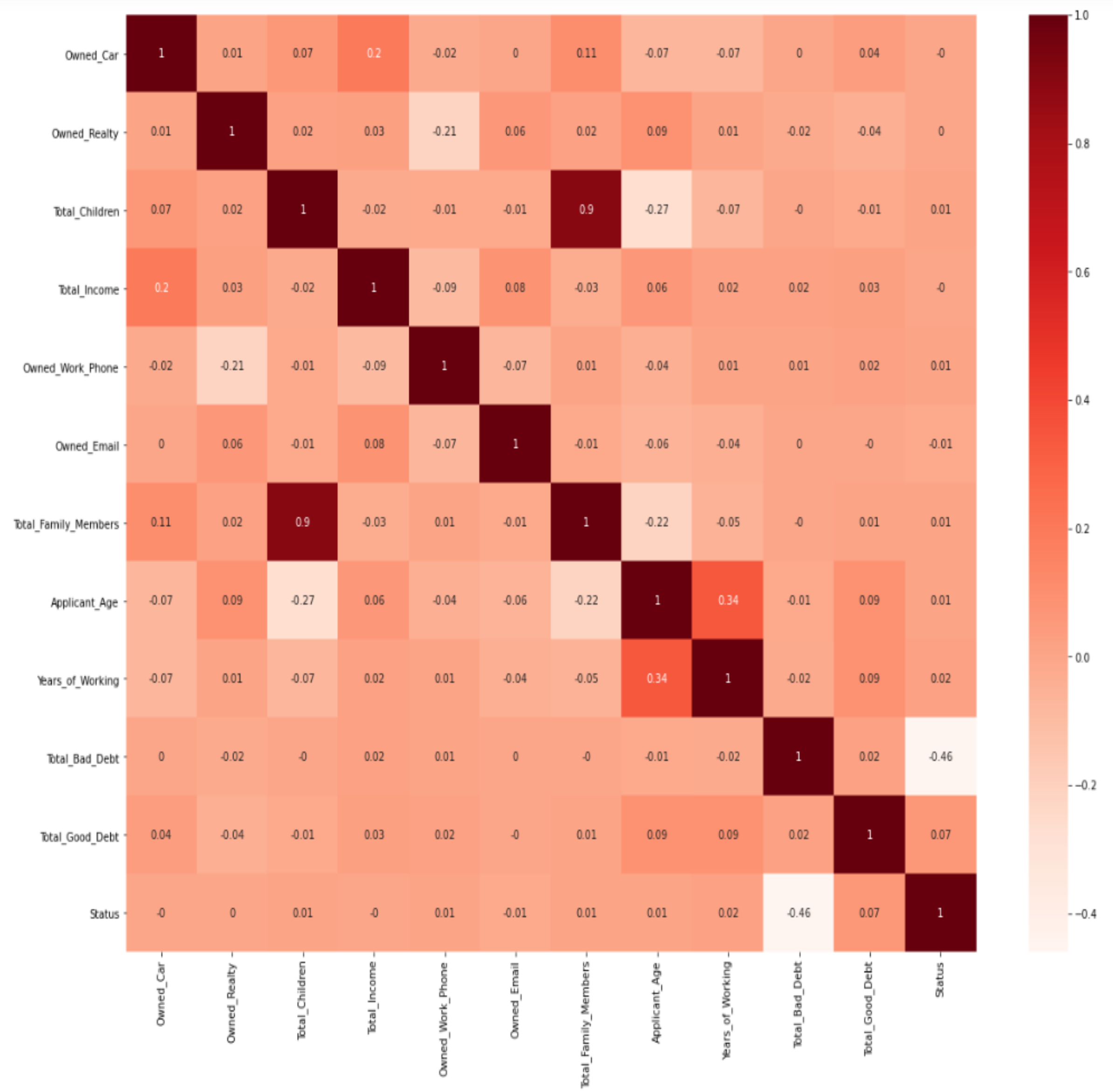
	A	B	C	D	E	F	G	H	I	J
1	Applicant_ID	Applicant_Gender	Owned_Car	Owned_Realty	Total_Children	Total_Income	Income_Type	Education_Type	Family_Status	Housing_Type
2	5008806	M	1	1	0	112500	Working	Secondary / secondary special	Married	House / apartment
3	5008808	F	0	1	0	270000	Commercial associate	Secondary / secondary special	Single / not married	House / apartment
4	5008809	F	0	1	0	270000	Commercial associate	Secondary / secondary special	Single / not married	House / apartment
5	5008810	F	0	1	0	270000	Commercial associate	Secondary / secondary special	Single / not married	House / apartment
6	5008811	F	0	1	0	270000	Commercial associate	Secondary / secondary special	Single / not married	House / apartment
7	5008815	M	1	1	0	270000	Working	Higher education	Married	House / apartment
8	5008819	M	1	1	0	135000	Commercial associate	Secondary / secondary special	Married	House / apartment
9	5008820	M	1	1	0	135000	Commercial associate	Secondary / secondary special	Married	House / apartment
10	5008821	M	1	1	0	135000	Commercial associate	Secondary / secondary special	Married	House / apartment
11	5008822	M	1	1	0	135000	Commercial associate	Secondary / secondary special	Married	House / apartment
12	5008823	M	1	1	0	135000	Commercial associate	Secondary / secondary special	Married	House / apartment
13	5008824	M	1	1	0	135000	Commercial associate	Secondary / secondary special	Married	House / apartment
14	5008825	F	1	0	0	130500	Working	Incomplete higher	Married	House / apartment
15	5008826	F	1	0	0	130500	Working	Incomplete higher	Married	House / apartment
16	5008830	F	0	1	0	157500	Working	Secondary / secondary special	Married	House / apartment
17	5008831	F	0	1	0	157500	Working	Secondary / secondary special	Married	House / apartment
18	5008832	F	0	1	0	157500	Working	Secondary / secondary special	Married	House / apartment
19	5008836	M	1	1	3	270000	Working	Secondary / secondary special	Married	House / apartment
20	5008837	M	1	1	3	270000	Working	Secondary / secondary special	Married	House / apartment
21	5008838	M	0	1	1	405000	Commercial associate	Higher education	Married	House / apartment
22	5008839	M	0	1	1	405000	Commercial associate	Higher education	Married	House / apartment
23	5008840	M	0	1	1	405000	Commercial associate	Higher education	Married	House / apartment
24	5008841	M	0	1	1	405000	Commercial associate	Higher education	Married	House / apartment
25	5008842	M	0	1	1	405000	Commercial associate	Higher education	Married	House / apartment



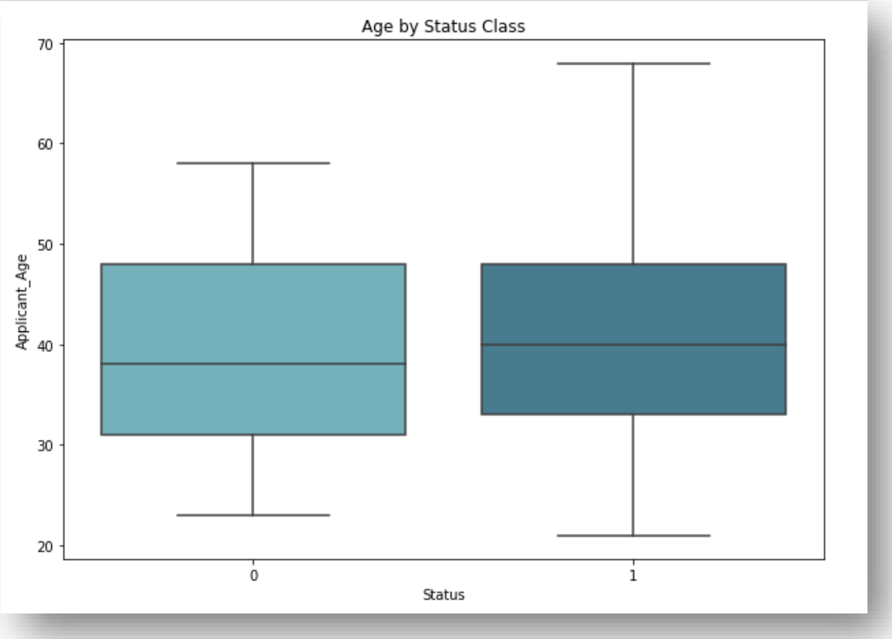
	K	L	M	N	O	P	Q	R	S	T	U
1	Owned_Mobile_Phone	Owned_Work_Phone	Owned_Phone	Owned_Email	Job_Title	Total_Family_Members	Applicant_Age	Years_of_Working	Total_Bad_Debt	Total_Good_Debt	Status
2	1	0	0	0	Security staff	2	59	4	0	30	1
3	1	0	1	1	Sales staff	1	53	9	0	5	1
4	1	0	1	1	Sales staff	1	53	9	0	5	1
5	1	0	1	1	Sales staff	1	53	9	0	27	1
6	1	0	1	1	Sales staff	1	53	9	0	39	1
7	1	1	1	1	Accountants	2	47	3	0	6	1
8	1	0	0	0	Laborers	2	49	4	0	8	1
9	1	0	0	0	Laborers	2	49	4	0	9	1
10	1	0	0	0	Laborers	2	49	4	0	9	1
11	1	0	0	0	Laborers	2	49	4	0	9	1
12	1	0	0	0	Laborers	2	49	4	0	5	1
13	1	0	0	0	Laborers	2	49	4	0	4	1
14	1	0	0	0	Accountants	2	30	4	1	25	1
15	1	0	0	0	Accountants	2	30	4	7	23	1
16	1	0	1	0	Laborers	2	28	5	2	30	1
17	1	0	1	0	Laborers	2	28	5	2	18	1
18	1	0	1	0	Laborers	2	28	5	2	33	1
19	1	0	0	0	Laborers	5	35	4	0	17	1
20	1	0	0	0	Laborers	5	35	4	0	17	1
21	1	0	0	0	Managers	3	33	6	0	31	1
22	1	0	0	0	Managers	3	33	6	0	14	1
23	1	0	0	0	Managers	3	33	6	0	56	1
24	1	0	0	0	Managers	3	33	6	0	5	1
25	1	0	0	0	Managers	3	33	6	0	9	1

3. Exploratory Data Analysis

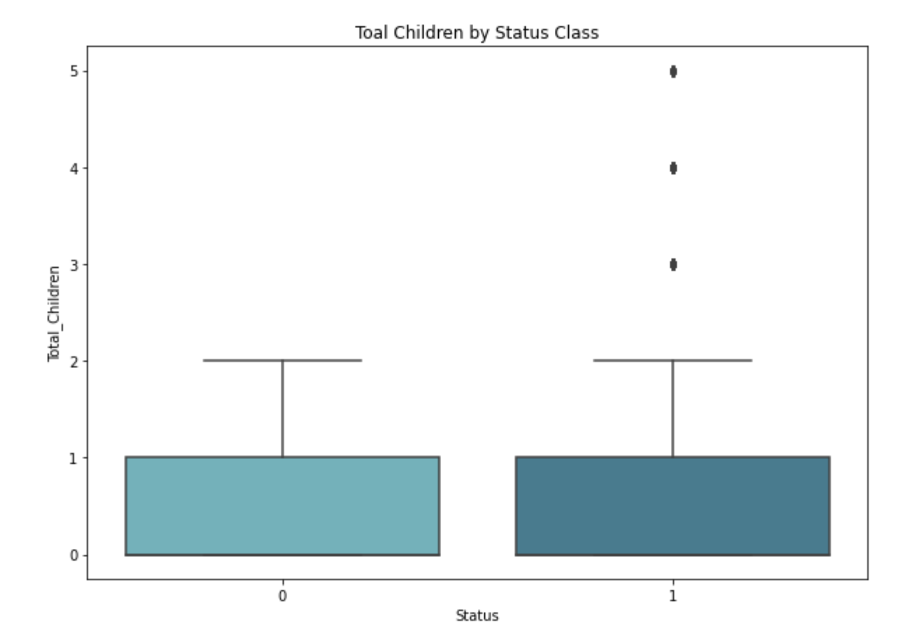
To check the relationship between the independent variables, we check the pairwise correlation between each independent variable. We do not see very high correlation between variables except total children and family size which is obvious. The other one is between the Applicant Age and Years of working as expected.



	Owned_Car	Owned_Realty	Total_Children	Total_Income	Owned_Work_Phone	Owned_Email	Total_Family_Members	Applicant_Age	Years_of_Working	Total_Bad_Debt	Total_Good_Debt	Status
Owned_Car	1	0.01	0.07	0.2	-0.02	0	0.11	-0.07	-0.07	0	0.04	-0
Owned_Realty	0.01	1	0.02	0.03	-0.21	0.06	0.02	0.09	0.01	-0.02	-0.04	0
Total_Children	0.07	0.02	1	-0.02	-0.01	-0.01	0.9	-0.27	-0.07	-0	-0.01	0.01
Total_Income	0.2	0.03	-0.02	1	-0.09	0.08	-0.03	0.06	0.02	0.02	0.03	-0
Owned_Work_Phone	-0.02	-0.21	-0.01	-0.09	1	-0.07	0.01	-0.04	0.01	0.01	0.02	0.01
Owned_Email	0	0.06	-0.01	0.08	-0.07	1	-0.01	-0.06	-0.04	0	-0	-0.01
Total_Family_Members	0.11	0.02	0.9	-0.03	0.01	-0.01	1	-0.22	-0.05	-0	0.01	0.01
Applicant_Age	-0.07	0.09	-0.27	0.06	-0.04	-0.06	-0.22	1	0.34	-0.01	0.09	0.01
Years_of_Working	-0.07	0.01	-0.07	0.02	0.01	-0.04	-0.05	0.34	1	-0.02	0.09	0.02
Total_Bad_Debt	0	-0.02	-0	0.02	0.01	0	-0	-0.01	-0.02	1	0.02	-0.46
Total_Good_Debt	0.04	-0.04	-0.01	0.03	0.02	-0	0.01	0.09	0.09	0.02	1	0.07
Status	-0	0	0.01	-0	0.01	-0.01	0.01	0.01	0.02	-0.46	0.07	1

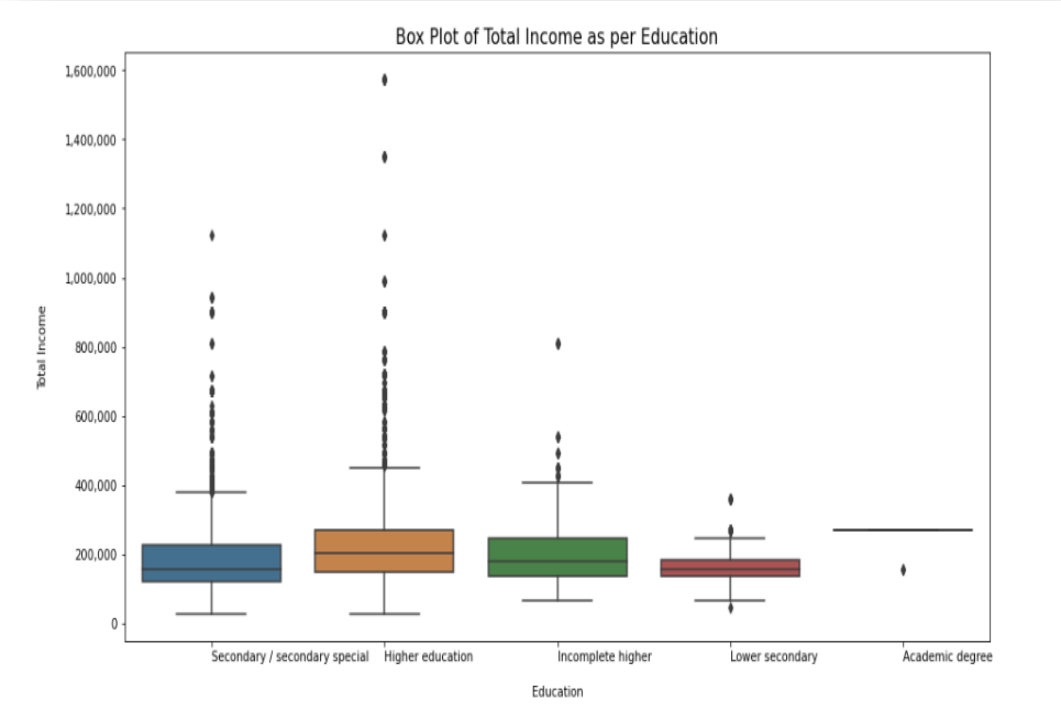


- We see that median age of approved and not approved applicants were almost same

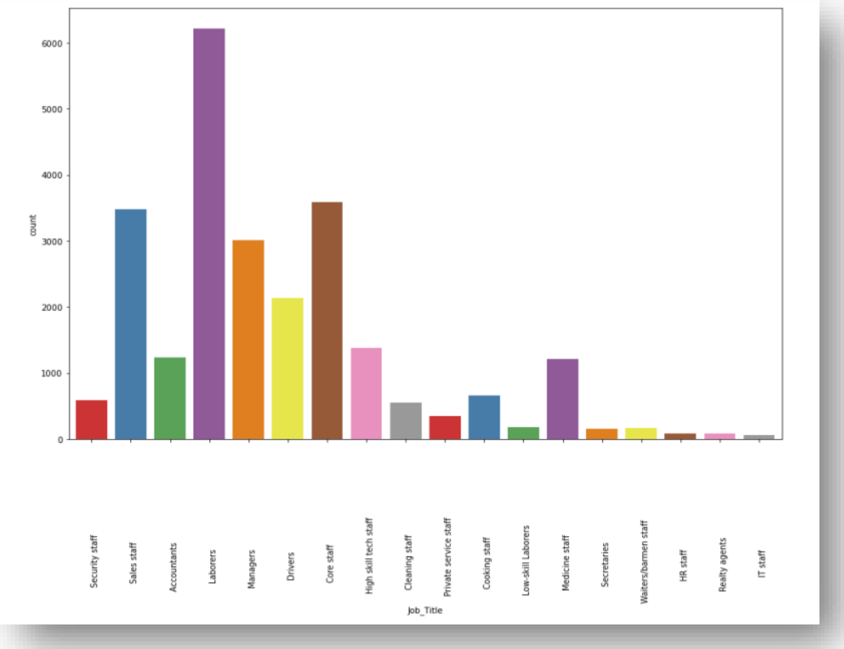


- Number of children for approved status has higher number of outliers but majority of applicants had 0 to 1 child

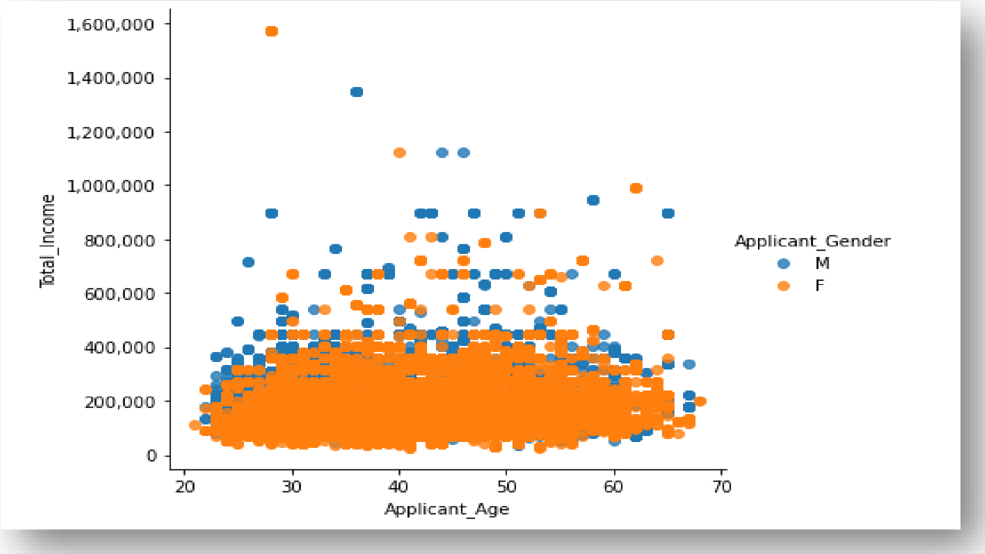
- We see a lot of outliers in all the Income type and median is almost similar. There are very few applicants with Academic degree
- We see a lot of outliers in all the Income type and median is almost similar.



- We checked the distribution of application basis profession.
- The bar graph shows that labor as a profession had the highest number of applicants for credit card approval followed by care staff and sales staff.



- Scatter plot of gender vs total income of applicants shows that in all applicants age range total income is consistent and has few outliers
- across all ages Female - 15627
- Male - 9501



4. Challenges:

For predictive modelling on banking dataset, we see that dependent variable is highly imbalanced. This can cause trouble in future machine learning models that will be used later; imbalanced data can cause a high bias problem, creating more Type 1 errors and/or Type 2 errors. With a very high imbalanced dataset model finds in difficult to ascertain the prediction and even though accuracy is high we can rely on the model because of the minority classes. Major challenge of imbalance dataset is that sometimes minority classes are useful, but machine learning algorithms are tended to be biased towards the majority classes and ignore the minority classes. To achieve the better result from our machine we need to train our machine using relevant data in such a way that makes those machines more talented and can give the accurate decision by itself for an unknown result.

Other challenge we faced was a greater number of categorical variables and with a greater number of levels in each variable. So, when we tried to encode categorical variables, our independent variables count increased heavily so we had to drop a lot categorical variables as well.

5. Feature Engineering:

In our data, the target variable that was whether credit card was approved or not was skewed, so we used SMOTE (Synthetic Minority Oversampling Technique) to synthetically oversample the data to remove bias.

SMOTE is an oversampling technique where synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

We applied SMOTE on training data and got 17501 approved and not approved applicants after oversampling and used it to train the models.

6. Performance Evaluation and Model Comparison:

Model	Test Accuracy	Train Accuracy	Precision	Recall	F1 Score	Support
Logistic Regression	99.93	99.87	0.87	1.00	0.93	33
Decision Tree	99.67	100	0.87	1.00	0.93	33
Random Forest	99.76	100	0.82	0.55	0.65	33
XGBoost	99.90	100	0.82	1.00	0.90	33
SVM	0.51	-	0.00	0.52	0.01	33

The accuracy of every model that we ran from Logistic Regression to XGBoost was about 99% which was expected because of the skewness of the dataset except for SVM which had a test accuracy of 51%.

Accuracy is not the right metric to compare in a dataset like ours as the decision variable is skewed, so we are considering precision, recall and f1-score as a metric to decide which model works best for our dataset.

We found that logistic regression model performs gives best result for this problem because of high accuracy and highest precision, recall and F-1 score.

7. Conclusion

We took on some of the most well-known pre-processing tasks, including scaling, label encoding, and missing value imputation, when developing this credit card approval prediction model. We used smote technique to tackle skewness of the data. We concluded with a machine learning model that could predict, given certain information about the applicant, whether their credit card application would be approved. This kind of machine learning model will be helpful for the banking and financial institution to reduce risk and time of approval hence enhancing customer experience and achieving objective and financial goals of the company.

8. Way Forward

This model is not only useful for banking systems but also be helpful to other financial institution in determining credit worthiness of a customer. We can take a step ahead and by making some changes in this model, converting this problem from classification to regression and instead of predicting approve and reject we can calculate risk associated with the applicant which will be a probability of a customer to not be a defaulter for a loan. Banks and financial institution can use this risk factor to decide the variable rate of interest and they can leverage it for subprime loans at higher rate of interest. This model can also be useful for various purposes like personal loan, auto loan, home loan etc.

9. References

Kaggle Link - https://www.kaggle.com/datasets/caesarmario/application-data
Smote: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

Thank You