CAPSTONE PROJECT

Pravin Ojha

OBJECTIVE AND DATA SOURCE

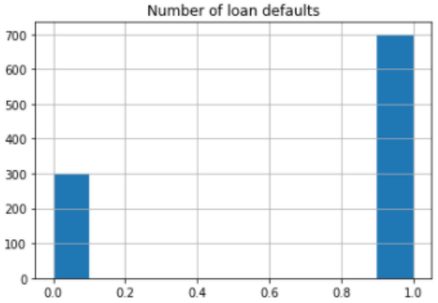
Context : Credit approval on unsecured loans is a complex process that requires banks to maintain high accuracy whilst deciding in shortest time-period

Objective : What are the determinants that can help predict credit default and thereby help to inform the credit approval process? Default being a binary outcome, intent is to build a model for in-principle digital lending decisioning

Approach :

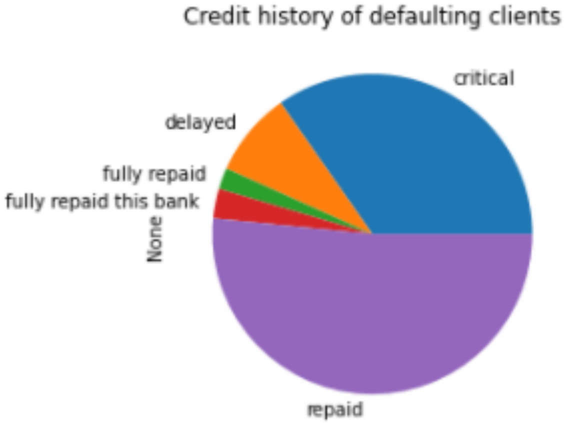
1. Credit approval Data was sourced from https://www.kaggle.com/shravan3273/credit-approval
2. EDA methods were used to understand the data
3. Cleansing and preparation of data was undertaken before Classifier models were run with focus on KNN, Log Reg and Decision-tree
4. Modeling techniques were used to improve model tuning and improving predication accuracy

OBSERVATIONS ON DATA

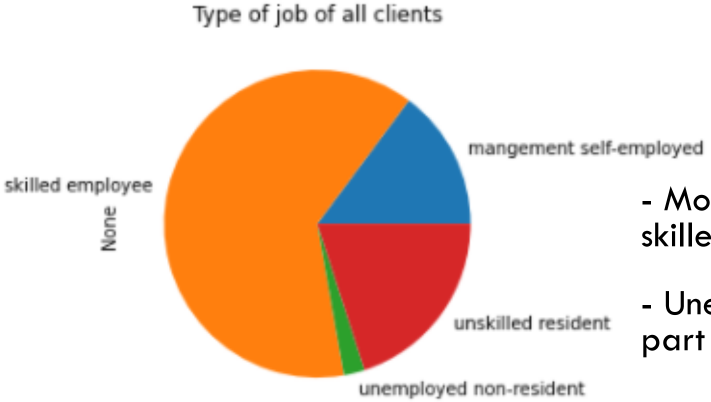


- 70% of data is loan default incident

- total 1000 data points x 21 column

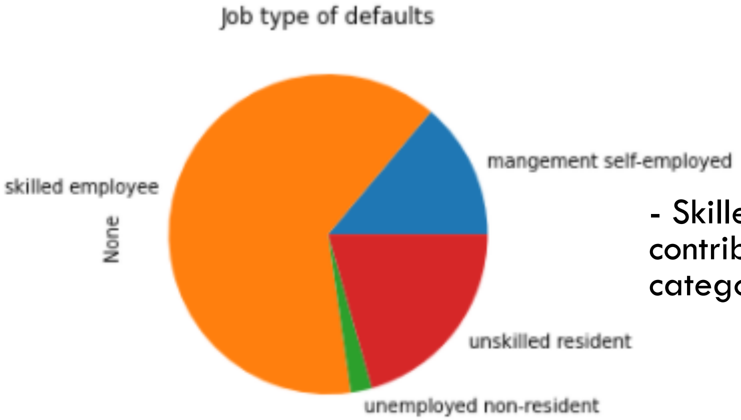


- cohort contains sizeable number with difficult credit history

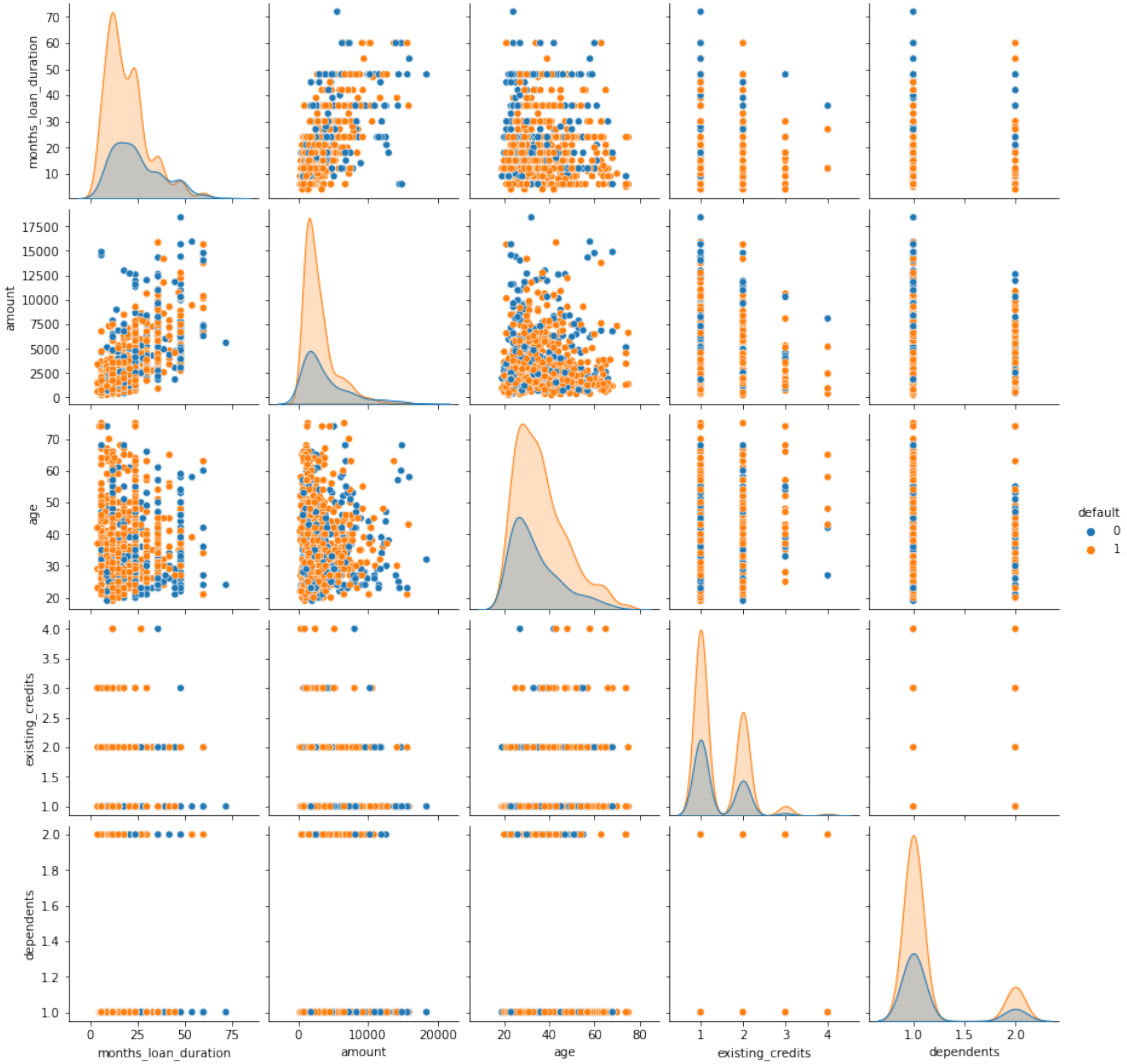


- More than 50% of the clients have skilled employee jobs

- Unemployed are not a significant part



- Skilled employees did contribute a lot to default category



OBSERVATIONS ON DATA

Combing through the above charts, some relationships seems to emerge. Higher likelihood of default with:

- higher months_loan_duration
- higher amount
- lower age (<35 yrs)
- existing_credits

~~Interesting to see no established relationship~~ with credit_histroy, employment_length and job.

OBSERVATIONS ON DATA

1. Stakeholders more likely to repay their loans without default:
   - skilled employees
   - management self-employees, divorced male and females followed by married male although small sample size
   - higher education and above
2. Customers those more likely to default on loans have
   - past credit history is bad (critical or delayed)
   - skilled employee
   - with employment tenure/length of less than 7 yrs with less than 4 yrs a bigger proportion
   - foreign workers
   - lesser dependents/family member
   - generally having lower savings balance
3. Loans taken for business, used cars, furniture, radio/tv are more likely to default

PREDICTIVE MODELS

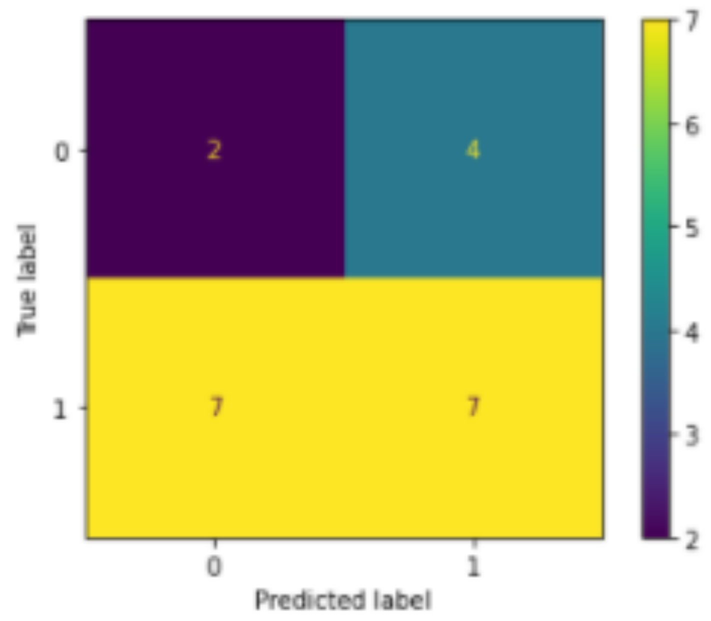
1. As the y is a prediction on ‘default’ following predictive models were trialed for suitability:
   - K Nearest Neighbors
   - Decision Leaf Tree and
   - Logistic Regression models were trialed to compare suitability of the models.
2. KNN and Decision-tree initial accuracy was in same range whist Logistic Regression was highly accurate
3. GridSearchCV was used to optimize the KNN and Decision tree models
4. KNN model significantly increased in accuracy

	Knearest Neighbors	**Decision Leaf Tree**	**Logistic Regression**
Initial accuracy	A = 0.628	A = 0.672	A = 1.0

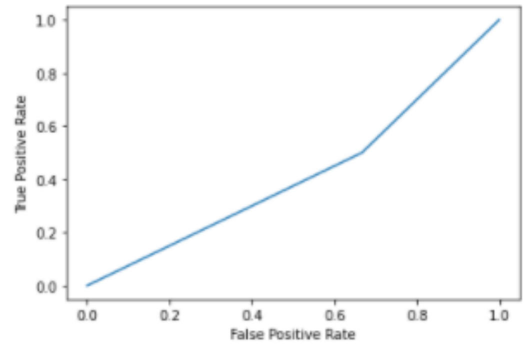
Optimised Accuracy	**A = 0.702**	**A = 0.7**	**NA**

MODEL SUITABILITY

1. Whilst accuracy improved upon optimization, inherently the model is not stable
2. AUC score of 0.41, Precision of 0.63 and Recall of 0.5 implies underlying weakness and instability in the model
3. False positives exist which means we predicted default, but it did not happen – means there is higher chance of credit not being approved even though client is eligible to get the credit - this is loss of revenue/interest incomes
4. False negative also exists which means we predicted no default but actually it was default - means still high chance of giving credit when the client is not worthy of receiving it. Which means credit delinquency costs are high
5. Overall, the model is not ready for customization and can be improved further – one way is to add more parameters and replace existing ones as well as increase sample size



```
AUC score:  0.41666666666666663
```


True Positive Rate
False Positive Rate