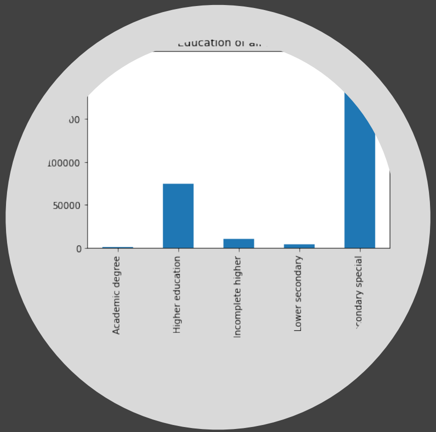
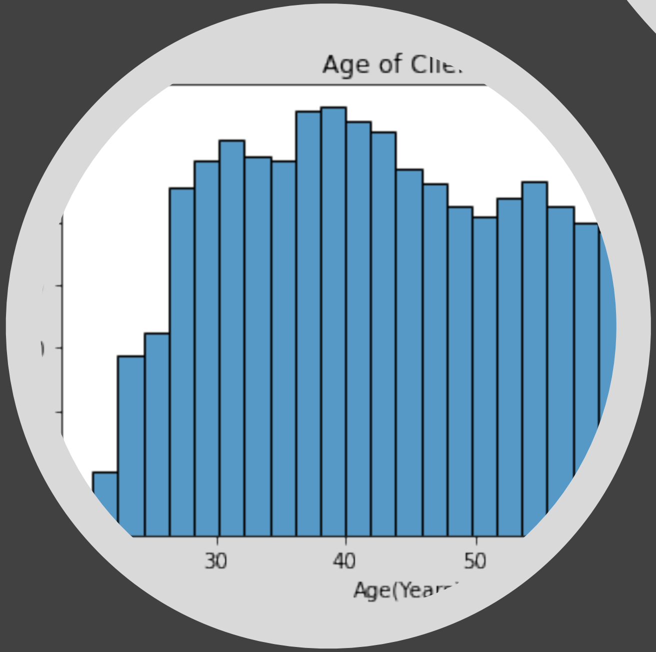
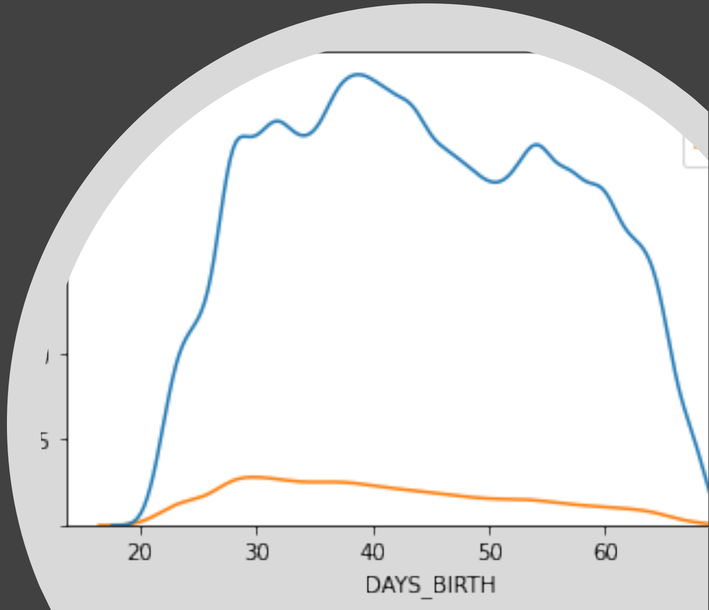
Can we find out whether a client has difficulty repaying their loan?

Overview of Data

- Data source: https://www.kaggle.com/c/home-credit-default-risk/overview
- Original data source (application train) was 307511 rows x 122 columns
    - Majority of the loans are paid on time, with 0.08% of loan defaulters.
    - Majority of the loan types are cash loans.
    - Several factors such as housing type, occupation type , income type, gender etc were analyzed

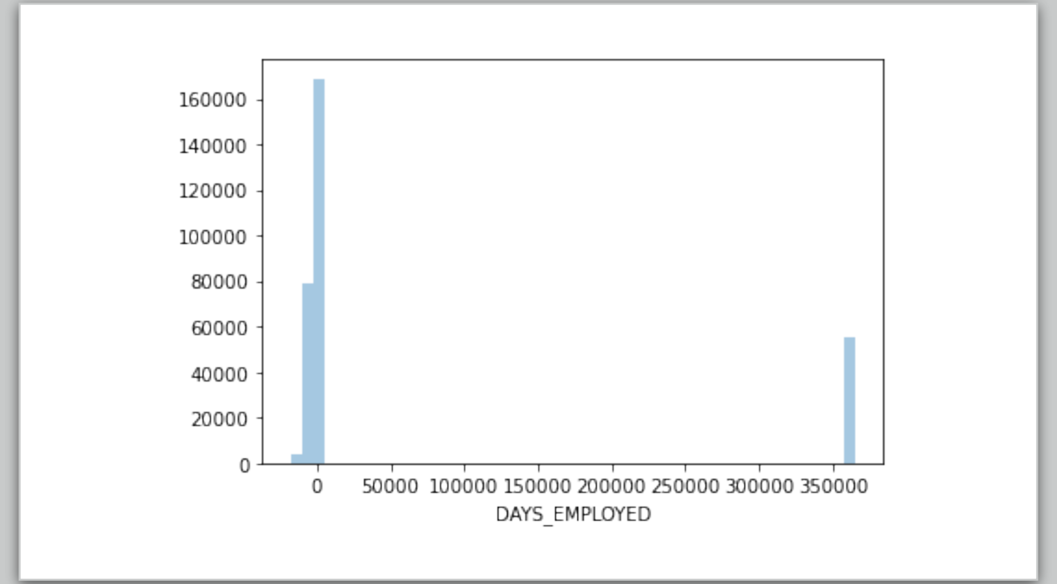


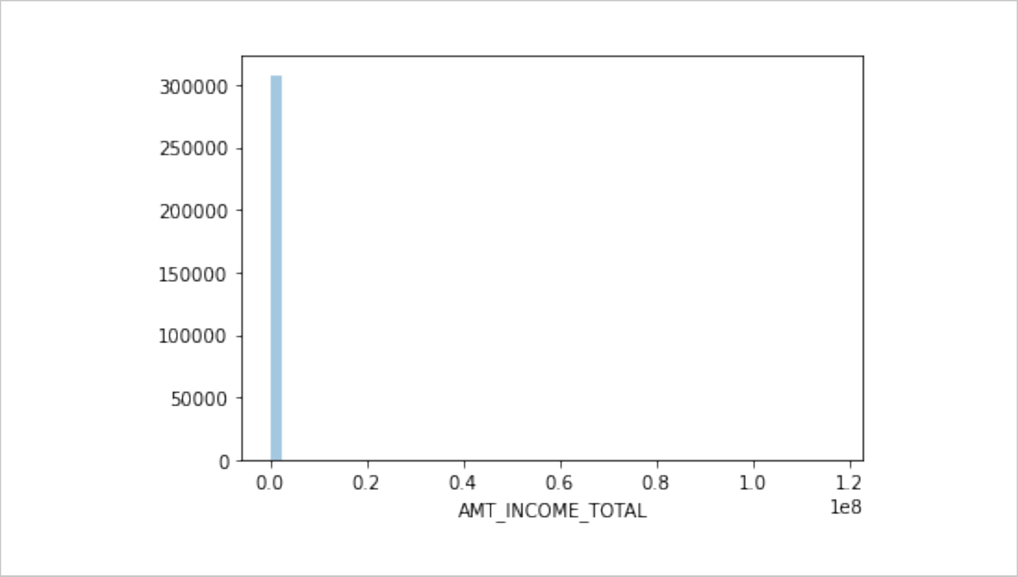




Data cleaning

- Data was checked for missing/duplicate values
- Outliers were handled
- Categorical data has to be encoded before model training because machine learning models deal with numbers only
- Feature scaling
- Data set has been split into training set and test set





Data Observations

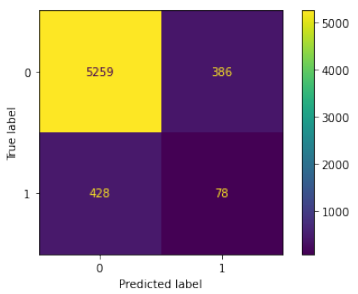
Higher likelihood of default

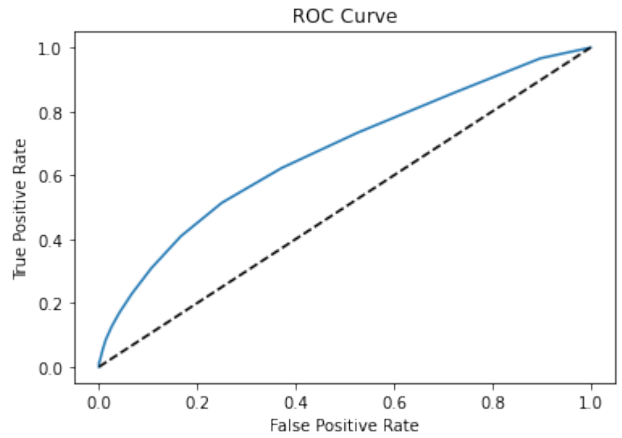
- Younger clients
- generally lower income, though there are anomalously high-end incomes
- Education level secondary school
- Mostly labourers

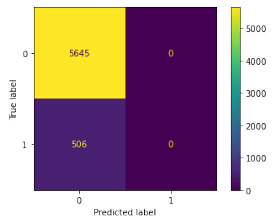
Predictive Models

KNearestNeighbors and LogisticRegression models were trialed to compare suitability of the models.

GridSearchCV was used to optimize the models







	LOGISTIC REGRESSION	KNN
ACCURACY	0.917	0.917

While the rate of accuracy is high, the ROC indicates that the predictions may be random.

It is likely so as model only predicted that ALL values would be 0 (would not default), so even though the predictor is 91.7%, it is far from accurate.

Given more time, these models could be improved with more variables included e.g. encoding categorical data to better train the model to predict.

- The KNN model and logistic reasoning both gave an accuracy of 91.7%. With greater time and expertise,i would try using the Decision Tree ML model. Furthermore, one thing i should have done is that i should have tried to use a scoped dataset instead.