MINI PROJECT 2: LOAN APPROVAL PREDICTION

Lillian Leong

Jun 2022

AGENDA

- Business Problem
- Overview of Dataset
- EDA & Summary of Dataset
- Evaluation of all Features vs Selected Features
- Summary of Models before/after Hyperparameter Tuning
- Conclusion

BUSINESS PROBLEM

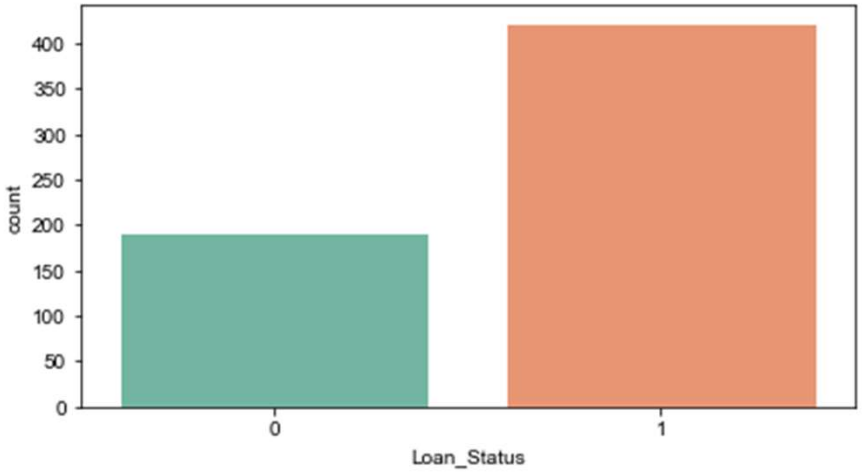
Our new online bank seeks to automate (in real time) the loan qualifying procedure based on information given by customers while filling out an online application form.

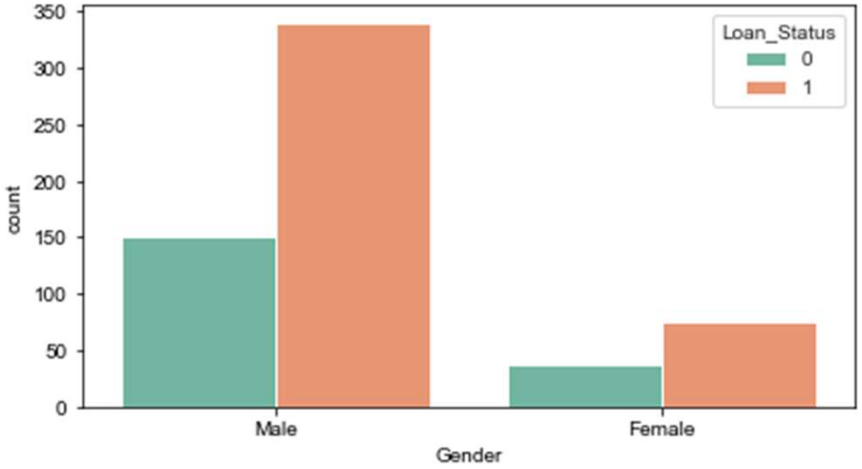
The data scientist team is approached to develop ML models that can help the company predict loan approval in accelerating decision-making process for determining whether an applicant is eligible for a loan or not.

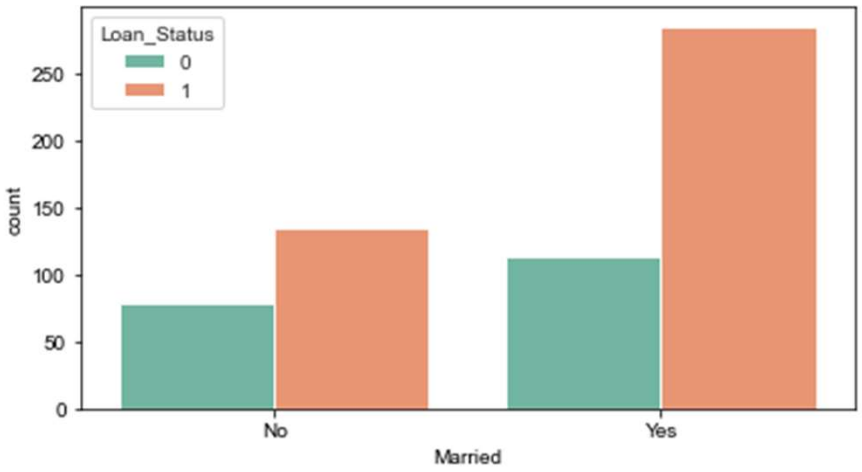
DATASET

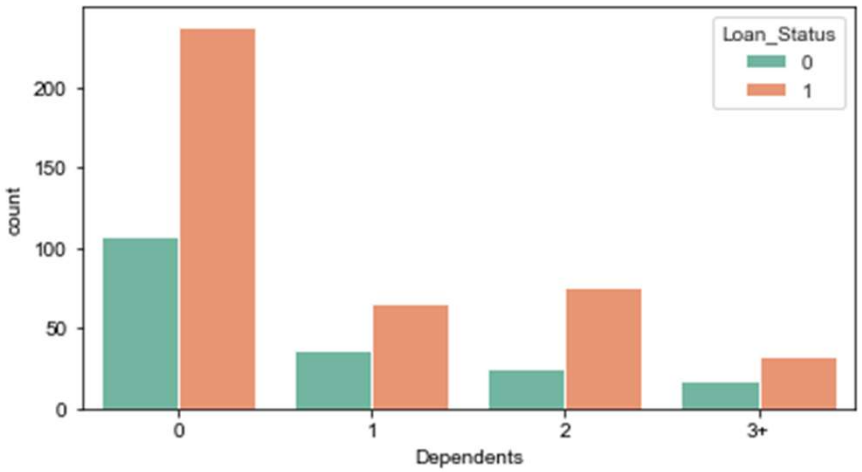
Col Name 614 records, 12 cols	Null Values	Data Preprocessing
Gender	13	5) Impute using the mode of 80% (male)
Married	3	6) Impute using the mode of 65% (married)
Dependents	15	4) Impute using the mode of 56% (0 dependent)
Education	0	
Self_Employed	32	4) Impute using the mode of 81% (not self employed)
ApplicantIncome	0	
CoapplicantIncome	0	
LoanAmount	22	1) Impute using ApplicantIncome
Loan_Amount_Term	14	2) Impute using the mode of 30 months (83%)
Credit_History	50	3) Impute using the mode of 30 months (77%)
Property_Area	0	
Loan_Status	0	

DATA VISUALISATION

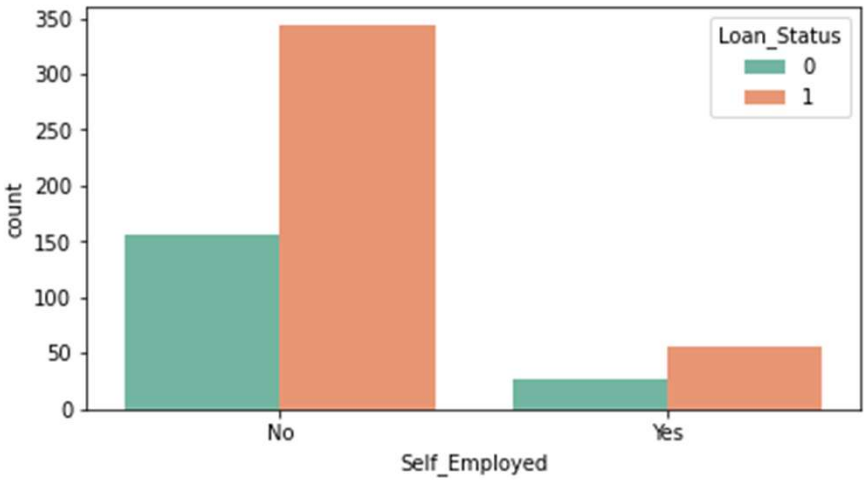


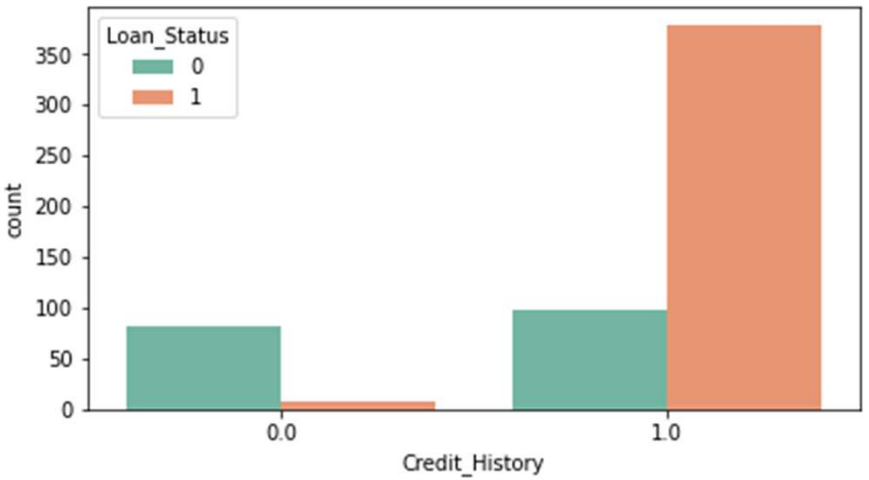


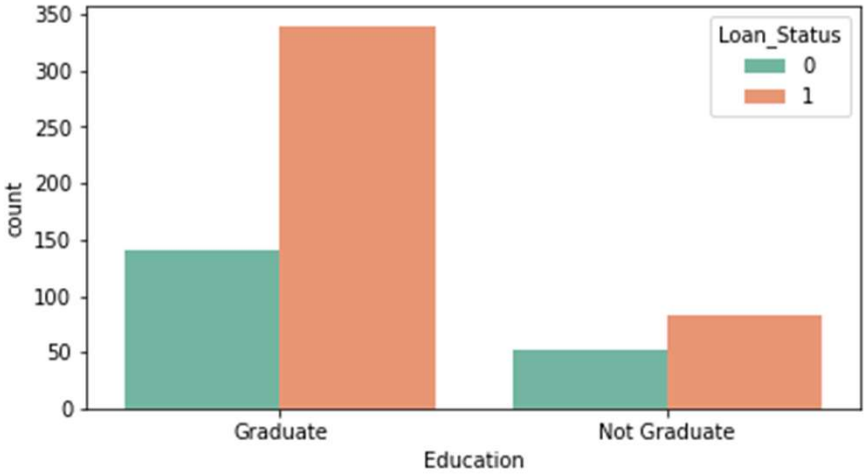


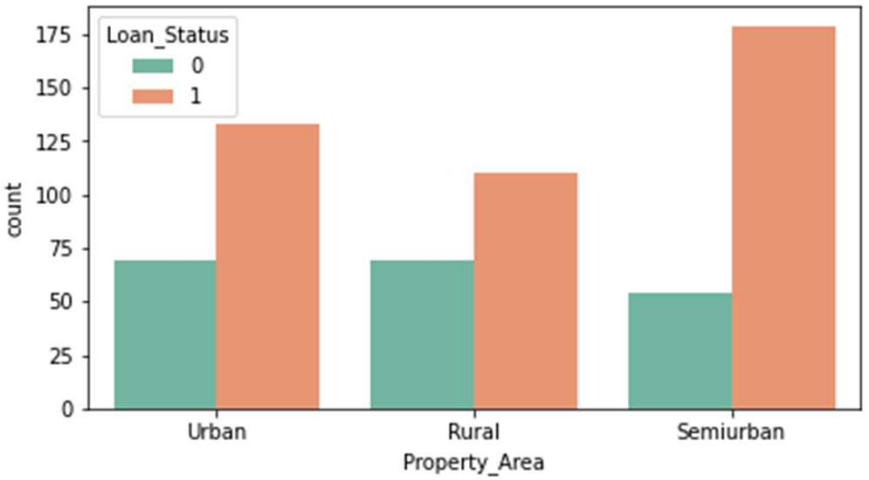


DATA VISUALISATION

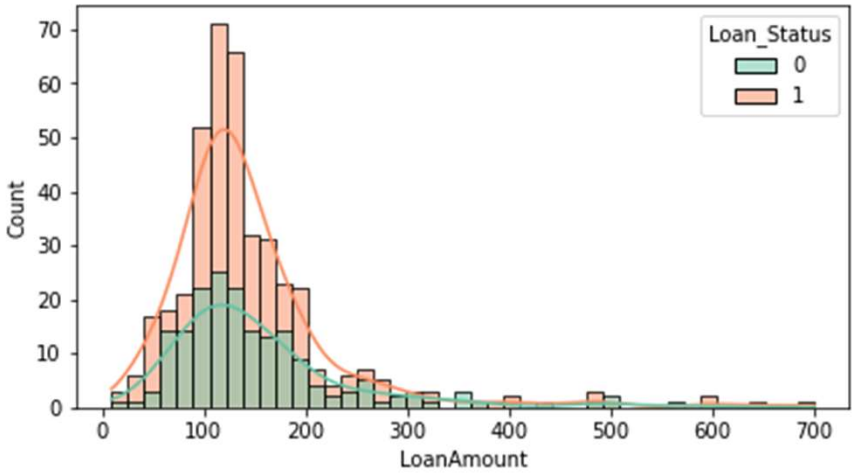


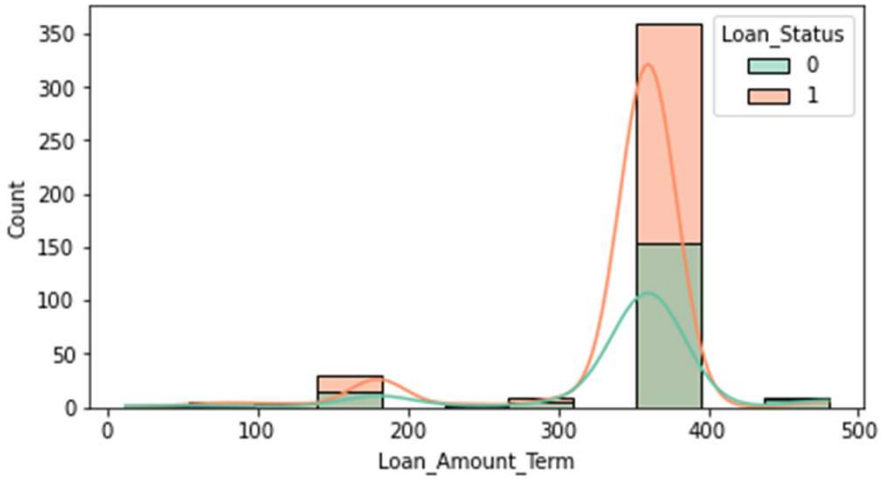


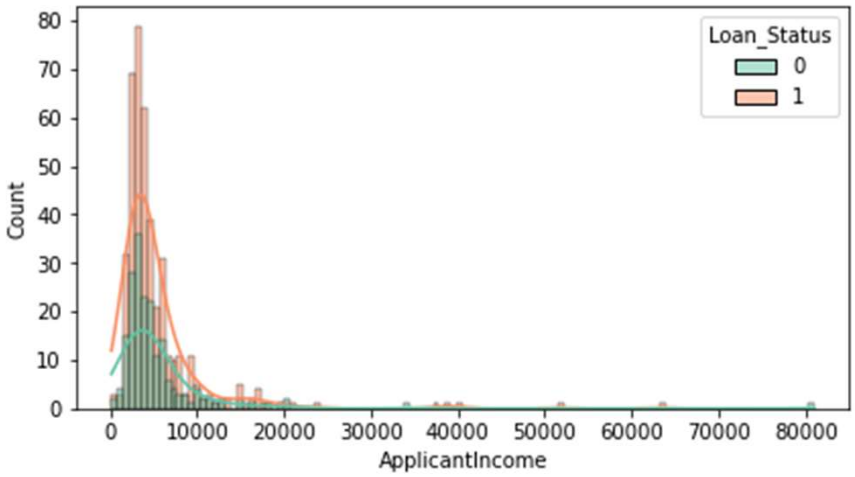


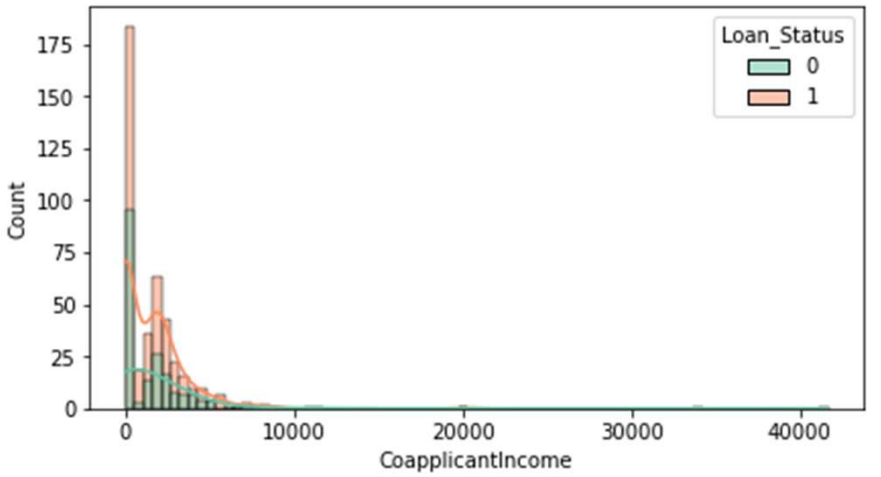


DATA VISUALISATION









SUMMARY OF DATASET

- There is higher proportion of loan status approved vs not approved.
- There is higher proportion of male applicants for approved loans.
- There is higher proportion of married applicants for approved loans.
- There is higher proportion of applicants with 0 dependents.
- There is higher proportion of non self employed in the approved loans
- There is a higher proportion of good credit history in the approved loans.
- There is a higher proportion of Graduate education in the approved loans.
- There is a higher proportion of Semiurban properties in the approved loans.

CORRELATION HEATMAP

Applied one hot encoding on selected columns

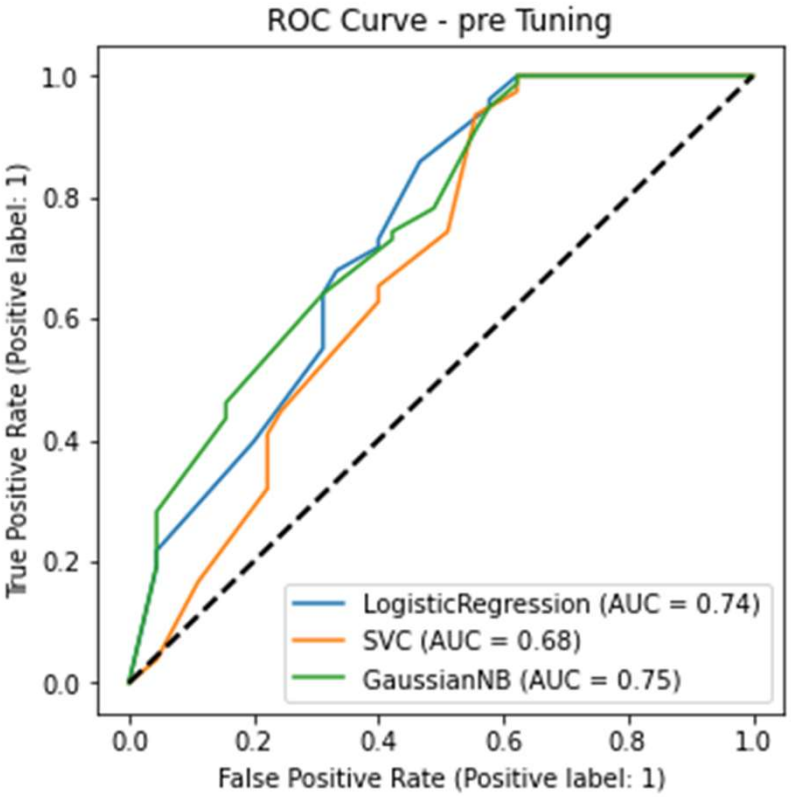
SUPERVISED LEARNING MODELS

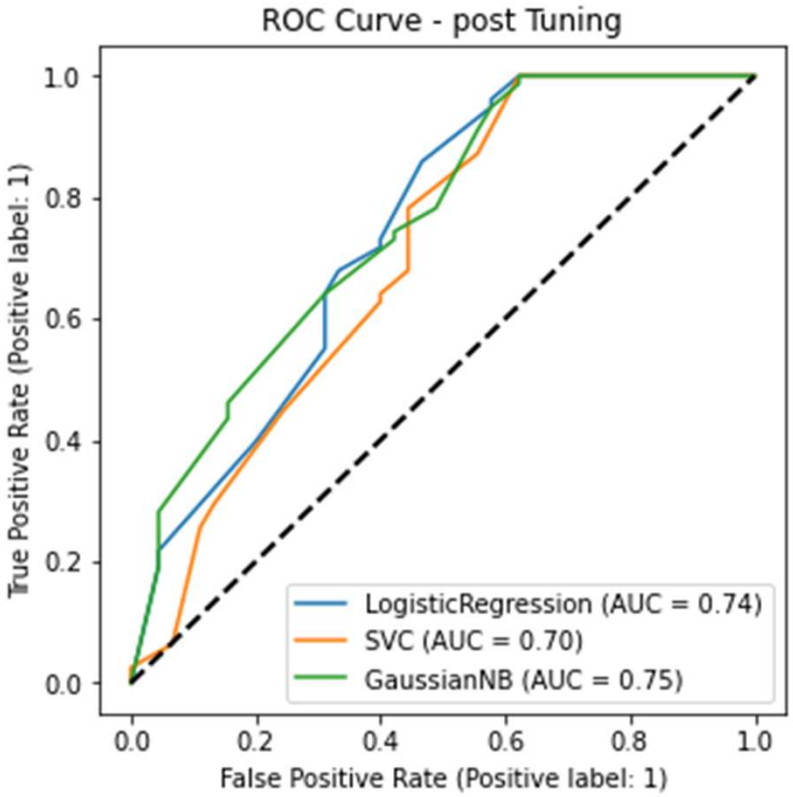
- The models used are
  - Logistic Regression
  - Support Vector Machine
  - Gaussian Naïve Bayes
- Recursive Feature Elimination
- Gridsearch CV to optimise the models

FEATURE SELECTION

- Feature ranking with recursive feature elimination
- Top 5 features selected
- Accuracy improved from 80.6% to 80.9%

ROC-AUC (PRE AND POST TUNING)





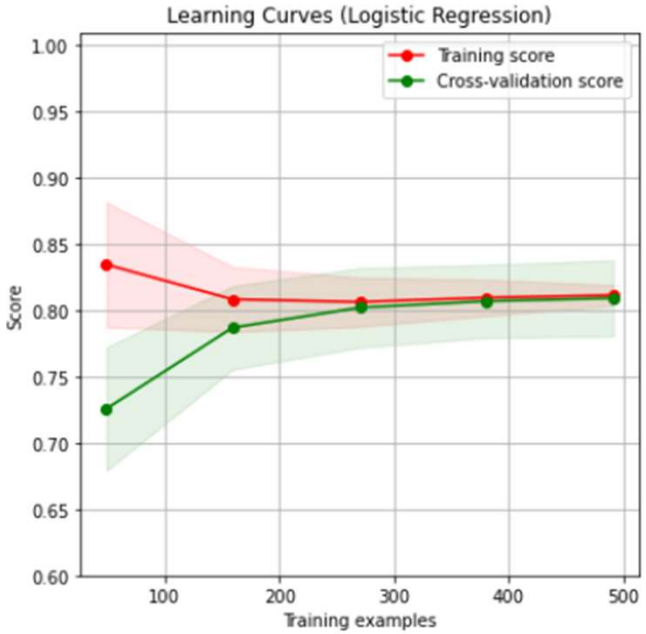
DETAIL METRICS

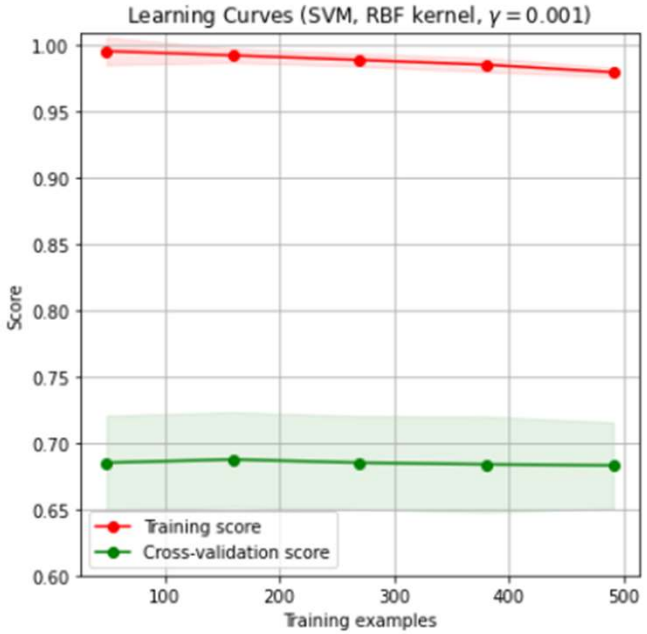
Model	Training Accuracy	Testing Accuracy	Precision	Recall	F1	AUC
After Hyperparameter Tuning						
Log Reg	0.8187	0.7724	0.7358	1.0	0.8478	0.7409
SVC	0.7006	0.6341	0.6341	1.0	0.7761	0.7027
Gaus NB	0.8187	0.7724	0.7358	1.0	0.8478	0.7548

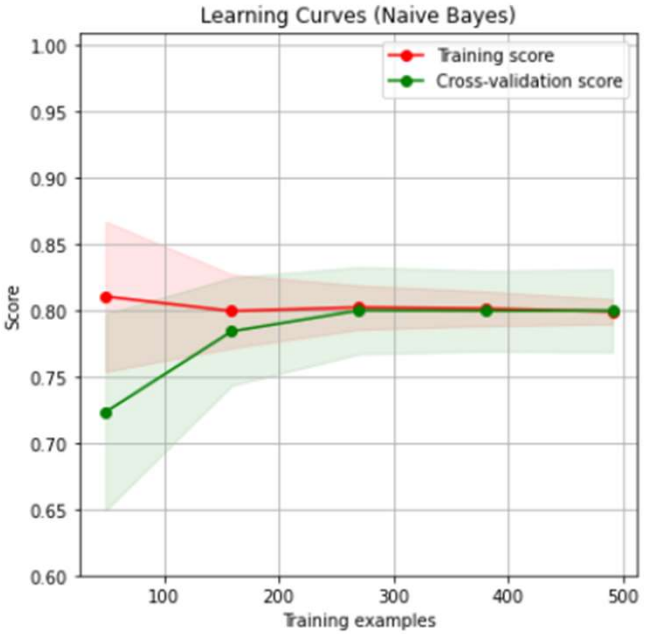
CONCLUSION & RECOMMENDATIONS

- Gaussian Naïve Bayes appears to be the most performing amongst the 3 models.
- Further investigation required on the dataset because different random state seems to have varying conclusion on the model performances.
- Can perform ensemble models on the datasets.

POST-TUNING LEARNING CURVES







POST-TUNING LEARNING CURVES