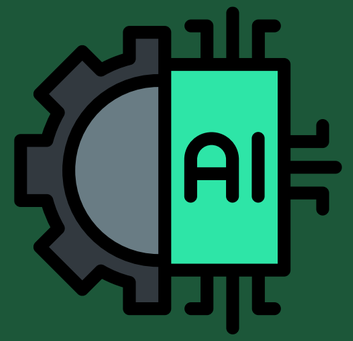
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

# A COMPARATIVE STUDY OF PREDICTING LOAN STATUS OF A LENDING COMPANY USING VARIOUS MACHINE LEARNING ALGORITHMS




Terrence Josiah
TP058242
APD3F2211ACS

# Content



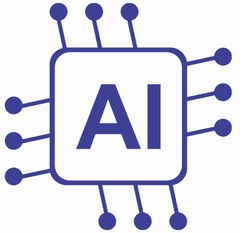
# ABSTRACT, INTRODUCTION & BACKGROUND


## Loan

- Increasing number of lending institutions
- Increasing demand for loan

## Data

- Data collection and availability
- The importance of data

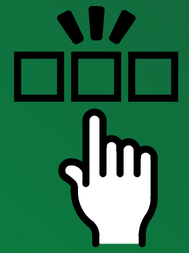



## Machine Learning

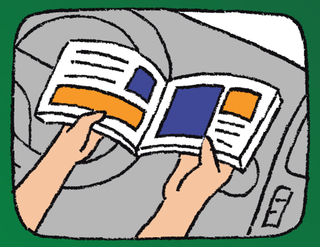
- The development of technology
- The use of Machine Learning

# Problem Statement

Risk of being solvent



Loan applicant selection



Traditional methods

Technology development

# Research Objectives & Questions

01

What are the machine learning models that are used to predict the loan grade?

02

From the machine learning models comparison, which model is performing the best?

03

What are the variables that influence loan grade the most?

*The objective of this research is to identify the best models to predict the risk associated with whether a borrower will be able to repay their loan on time, which is reflected by the loan grade, as well as to find the most influential variables towards the loan grade*

# Significance/Scope/Limitations

## Significance

- Solving various financial sector issues
- The needs of highly accurate predictive modelling system
- Mitigating risks

## Scope

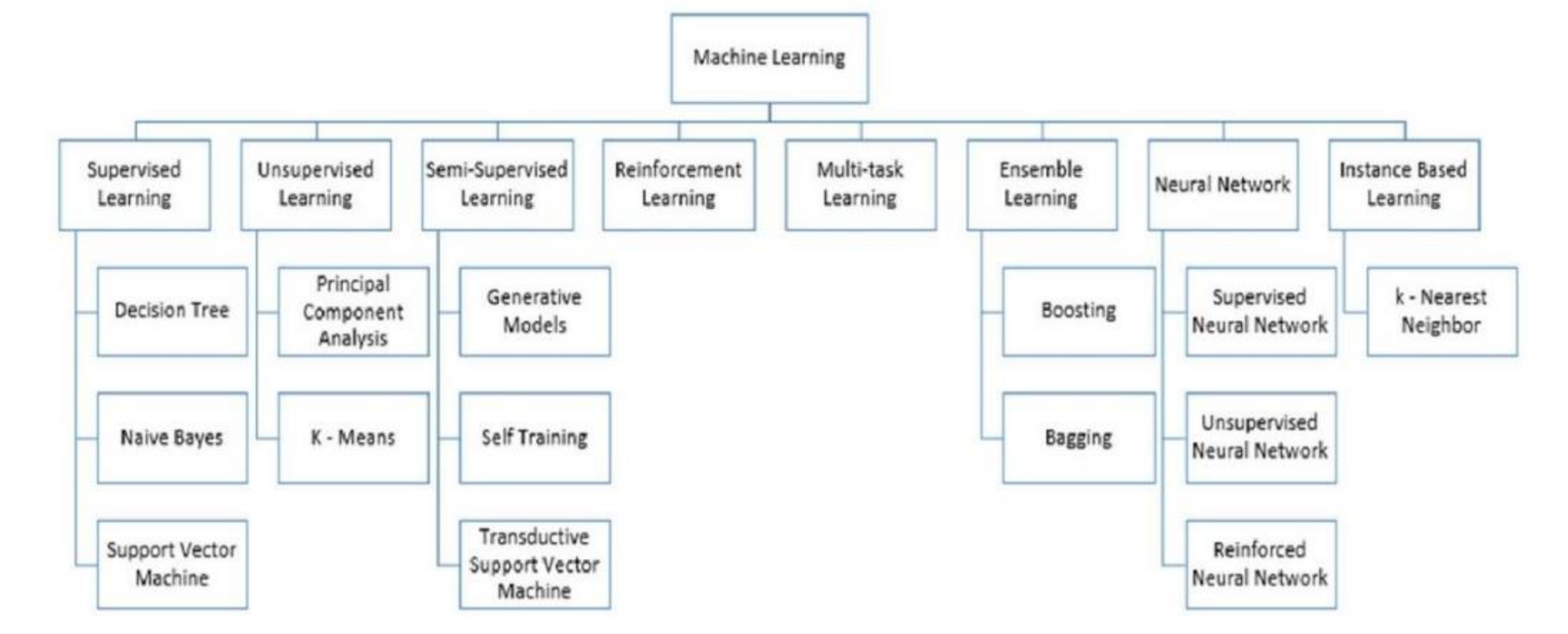
- Purpose: identify the best machine learning models in predicting the loan grade
- The scope is defined by the dependent variables, location, and the targets (borrowers)
- Time scope: 2007 - 2018
- Data collection: from LendingClub officia website

## Limitations

- Selection bias: data only from one company
- Time and resource constraints
- Device limitations
- non-optimized number of samples

# LITERATURE REVIEW

# Machine Learning Algorithms



# Defining Loan

The act of lending something, between parties, with a set of conditions to be met.

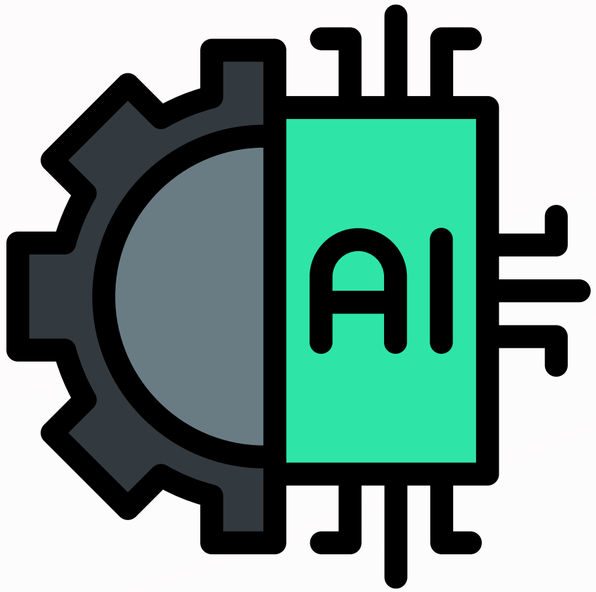
Plays a crucial role in providing opportunities

Secured Loan

Unsecured Loan

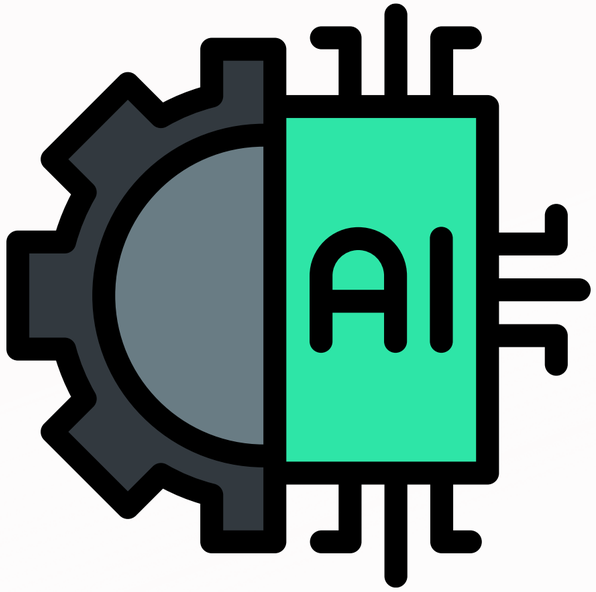
Close-End & Open-End Loan

# Development of Machine Learning Algorithms



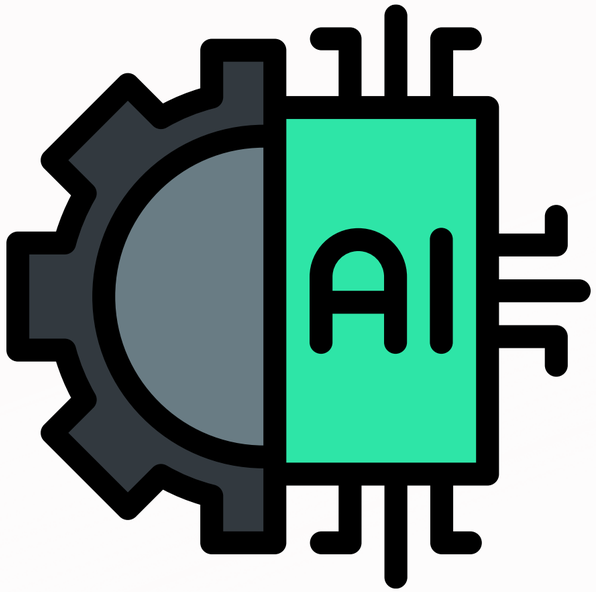

| | |
|---|---|
| 1950 | Few of the first development called the "Turning Test" was founded by Alan Turning with the purpose of checking a machine's intelligence. The "pass" requirement can be achieved if the machine is able to talk to real human without the human thinking that they are actually talking to machine. |
| 1952 | A highly adaptive algorithm was created by Samuel. It has the ability to play the Game of Checkers and train as well as adapt to various kinds of possibilities by itself. |
| 1956 | Artificial Intelligence was born from the meeting of 4 researchers in 1956, notably in DartMouth. |
| 1958 | The creation of Perceptron as a foundation in developing Artificial Neural Network (ANN) by Frank Rosenblatt. |
| 1967 | The proposal of the Nearest Neighbor Algorithm was issued, which can be used for recognizing patterns. |
| 1979 | Standford Cart, which is basically a robot could move through a space and avoid objects in its way was created and developed by some students in Standford University. |

# Development of Machine Learning Algorithms




| | |
|---|---|
| 1981 | Gerald Dejong made a proposal regarding the Explanation Based Learning (EBL). Additionally, the training data can be analysed by a computer to develop rules for eliminating unnecessary data. |
| 1985 | The invention of NetTalk, which is an algorithm that was able to learn how to pronounce words in English similar to the methods that children learn, was realized by Terry Sejnowski. |
| 1990 | The shifting of the main objective of Machine Learning from Knowledge-Driven to Data-Driven occurred. Machine Learning at this point was primarily used to perform analysis of large datasets as well as deriving conclusions from it. |
| 1997 | Gary Kasparov, which is the World Chess Champion, was defeated by Deep Blue computer developed by IBM. |
| 2006 | Geoffery Hinton first popularised the phrase "Deep Learning" to describe a novel neural network architecture that utilised numerous layers of neurons to facilitate learning. |
| 2011 | An algorithm-based IBM's Watson was successfully defeated a real human at Jeopardy Game. IBM's Watson was able to respond to inquiries in natural language. |

# Development of Machine Learning Algorithms




| | |
|---|---|
| 2012 | The detection of patterns in videos and image was made possible through the invention of GoogleBrain by Jeff Dean. |
| 2014 | The identification of human faces was made possible through the invention of DeepFace by Facebook. |
| 2015 | Amazon suggested a machine learning platform of its own. The "Distributed Machine Learning Toolkit" was developed by Microsoft to facilitate the effective distribution of machine learning issues to numerous machines for parallel computation to discover a solution. With the aim of employing artificial intelligence to benefit humans, Elon Musk and Sam Altman founded the non-profit foundation OoeaAI. |
| 2016 | The most complicated board game, DeepMind, was proposed by Google. The Google AlphaGo software beats a skilled human player in go for the first time. It is based on combining approaches from tree searching and machine learning. |
| 2017 | Google proposed the machine learning and deep learning-based Google Lens, Google Clicks, Google Home Mini, and Google Nexus phones. The Deep Learning Engine: NVIDIA GPUs was a proposal made by Nvidia. Home Pod, an interactive machine learning gadget, was proposed by Apple. |

# Sampling & Data Collection

**SOURCE:**

**LENDINGCLUB**
**OFFICIAL WEBSITE**

**SECONDARY DATA**
**2007 - 2018**

LendingClub

# DATA ANALYSIS

# Missing Values

| | | |
|---|---|---|
| sec_app_earliest_cr_line | 2152680 | 95.20 |
| sec_app_inq_last_6mths | 2152680 | 95.20 |
| sec_app_num_rev_accts | 2152680 | 95.20 |
| sec_app_open_act_il | 2152680 | 95.20 |
| sec_app_open_acc | 2152680 | 95.20 |
| verification_status_joint | 2144971 | 94.90 |
| dti_joint | 2139995 | 94.70 |
| annual_inc_joint | 2139991 | 94.70 |
| desc | 2134636 | 94.40 |
| mths_since_last_record | 1901545 | 84.10 |
| mths_since_recent_bc_dlq | 1741000 | 77.00 |
| mths_since_last_major_derog | 1679926 | 74.30 |

Your selected dataframe has 151 columns.
There are 150 columns that have missing values.

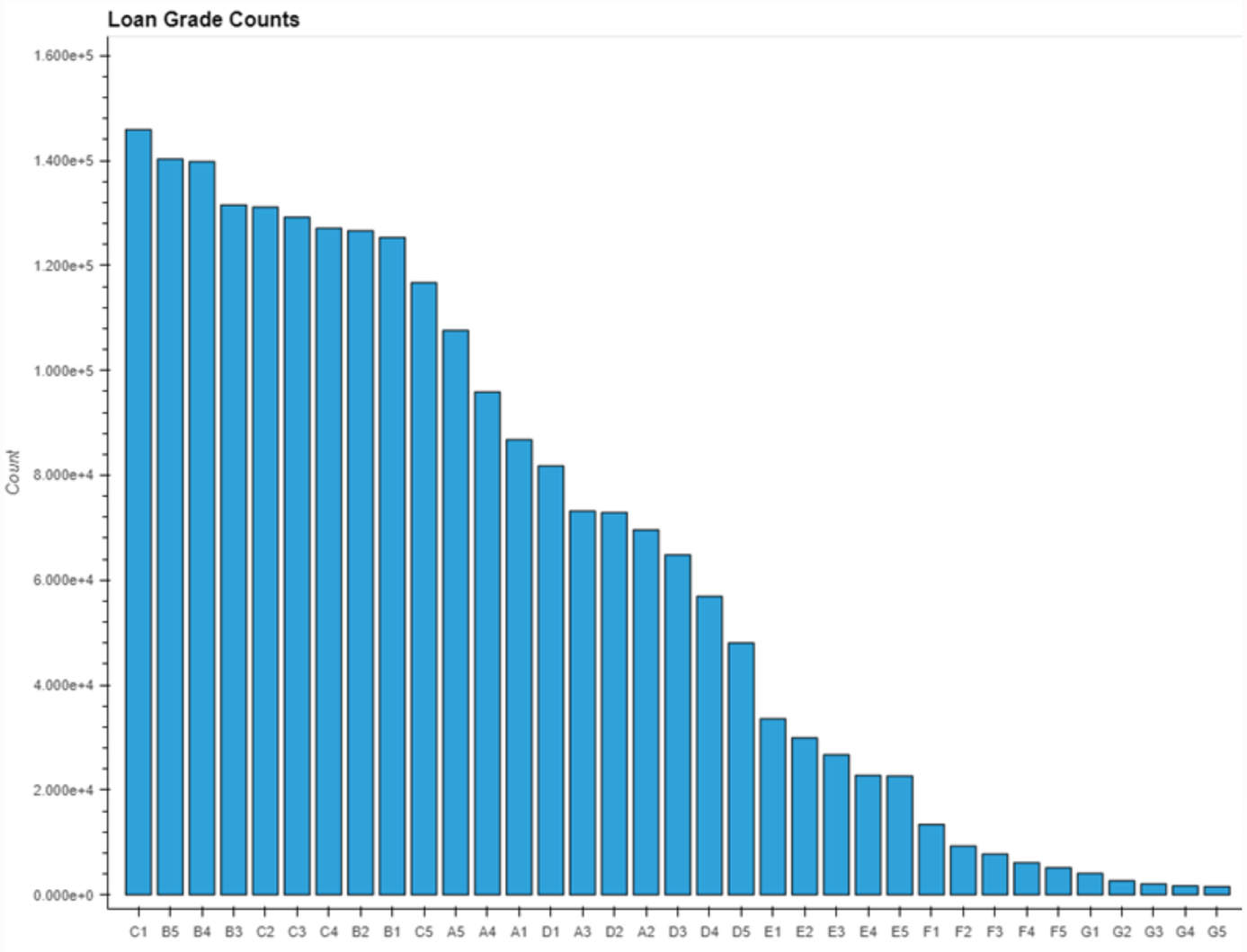
| | Missing Values | % of Total Values |
|---|---|---|
| member_id | 2260701 | 100.00 |
| orig_projected_additional_accrued_interest | 2252050 | 99.60 |
| hardship_dpd | 2249784 | 99.50 |
| hardship_status | 2249784 | 99.50 |
| deferral_term | 2249784 | 99.50 |
| hardship_amount | 2249784 | 99.50 |
| hardship_start_date | 2249784 | 99.50 |
| hardship_end_date | 2249784 | 99.50 |
| payment_plan_start_date | 2249784 | 99.50 |
| hardship_length | 2249784 | 99.50 |
| hardship_loan_status | 2249784 | 99.50 |
| hardship_type | 2249784 | 99.50 |
| hardship_payoff_balance_amount | 2249784 | 99.50 |
| hardship_last_payment_amount | 2249784 | 99.50 |
| hardship_reason | 2249784 | 99.50 |
| debt_settlement_flag_date | 2226455 | 98.50 |
| settlement_status | 2226455 | 98.50 |
| settlement_date | 2226455 | 98.50 |
| settlement_amount | 2226455 | 98.50 |
| settlement_percentage | 2226455 | 98.50 |
| settlement_term | 2226455 | 98.50 |
| sec_app_mths_since_last_major_derog | 2224759 | 98.40 |
| sec_app_revol_util | 2154517 | 95.30 |
| revol_bal_joint | 2152681 | 95.20 |
| sec_app_fico_range_high | 2152680 | 95.20 |
| sec_app_mort_acc | 2152680 | 95.20 |
| sec_app_fico_range_low | 2152680 | 95.20 |
| sec_app_collections_12_mths_ex_med | 2152680 | 95.20 |
| sec_app_chargeoff_within_12_mths | 2152680 | 95.20 |

# Correlation Plot

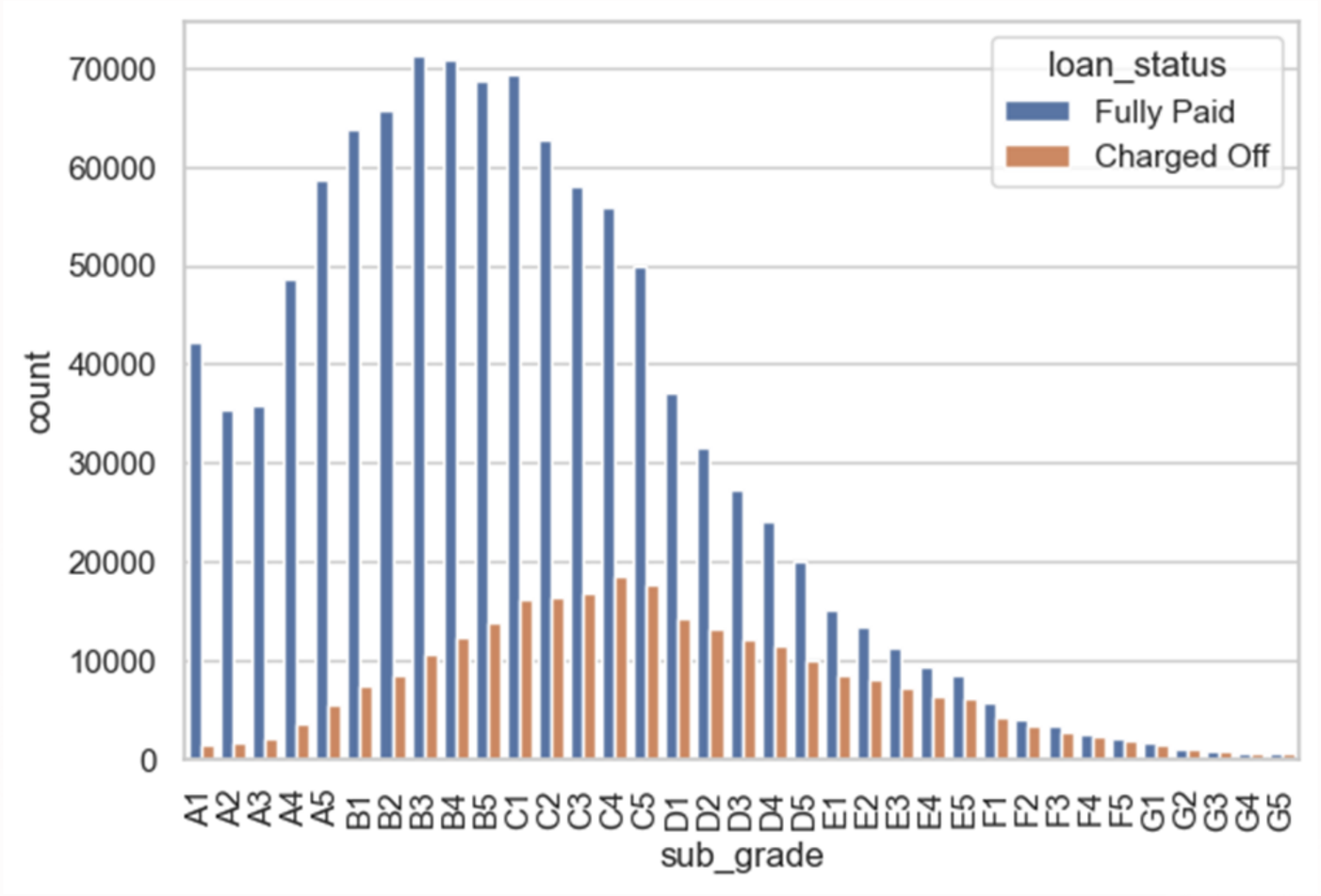
## Alerts

| | |
|---|---|
| `loan_amnt` is highly overall correlated with `installment` | High correlation |
| `int_rate` is highly overall correlated with `sub_grade` | High correlation |
| `installment` is highly overall correlated with `loan_amnt` | High correlation |
| `open_acc` is highly overall correlated with `total_acc` | High correlation |
| `pub_rec` is highly overall correlated with `pub_rec_bankruptcies` | High correlation |
| `revol_bal` is highly overall correlated with `revol_util` | High correlation |
| `revol_util` is highly overall correlated with `revol_bal` | High correlation |
| `total_acc` is highly overall correlated with `open_acc` | High correlation |
| `pub_rec_bankruptcies` is highly overall correlated with `pub_rec` | High correlation |
| `sub_grade` is highly overall correlated with `int_rate` | High correlation |

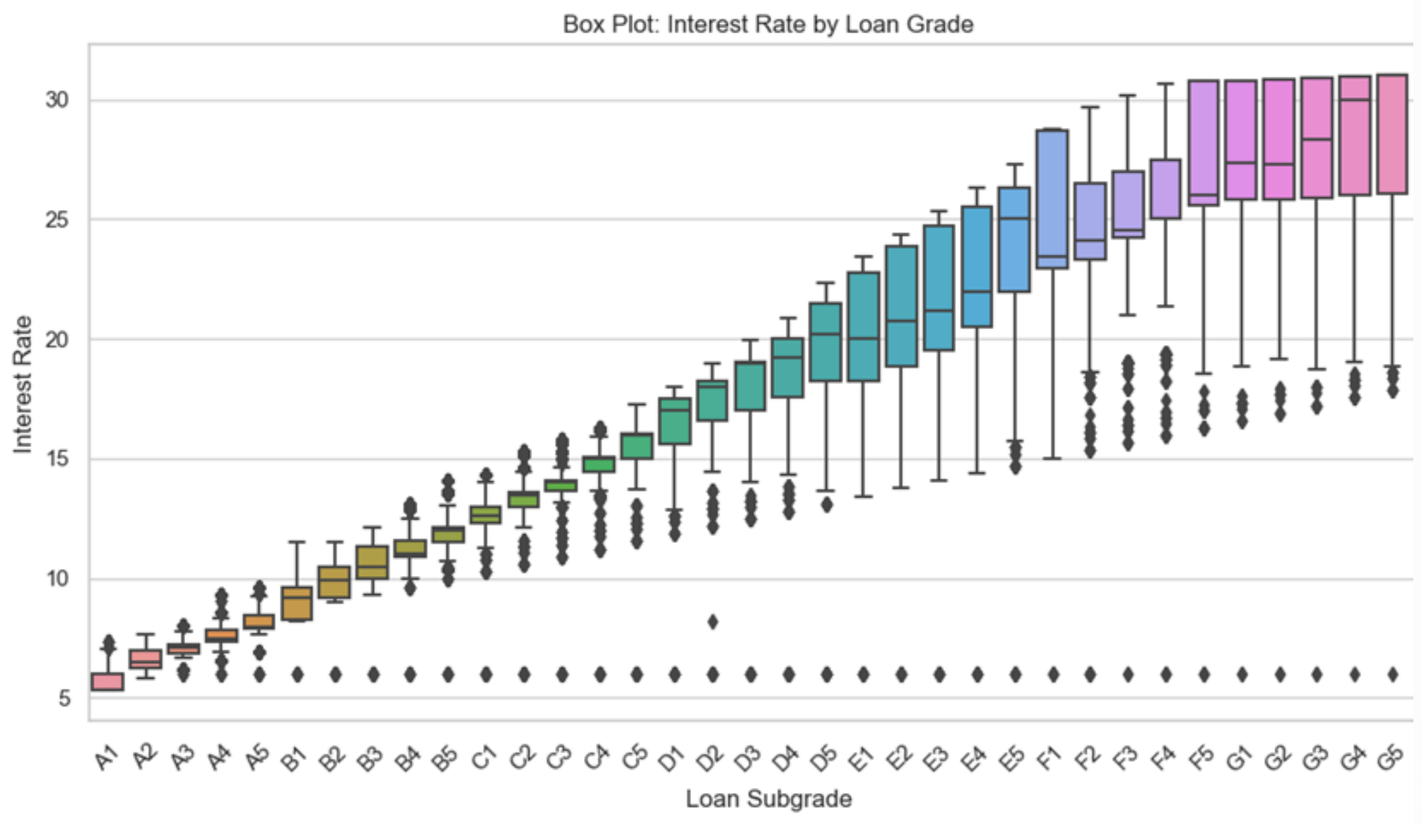
# Loan Grade Counts



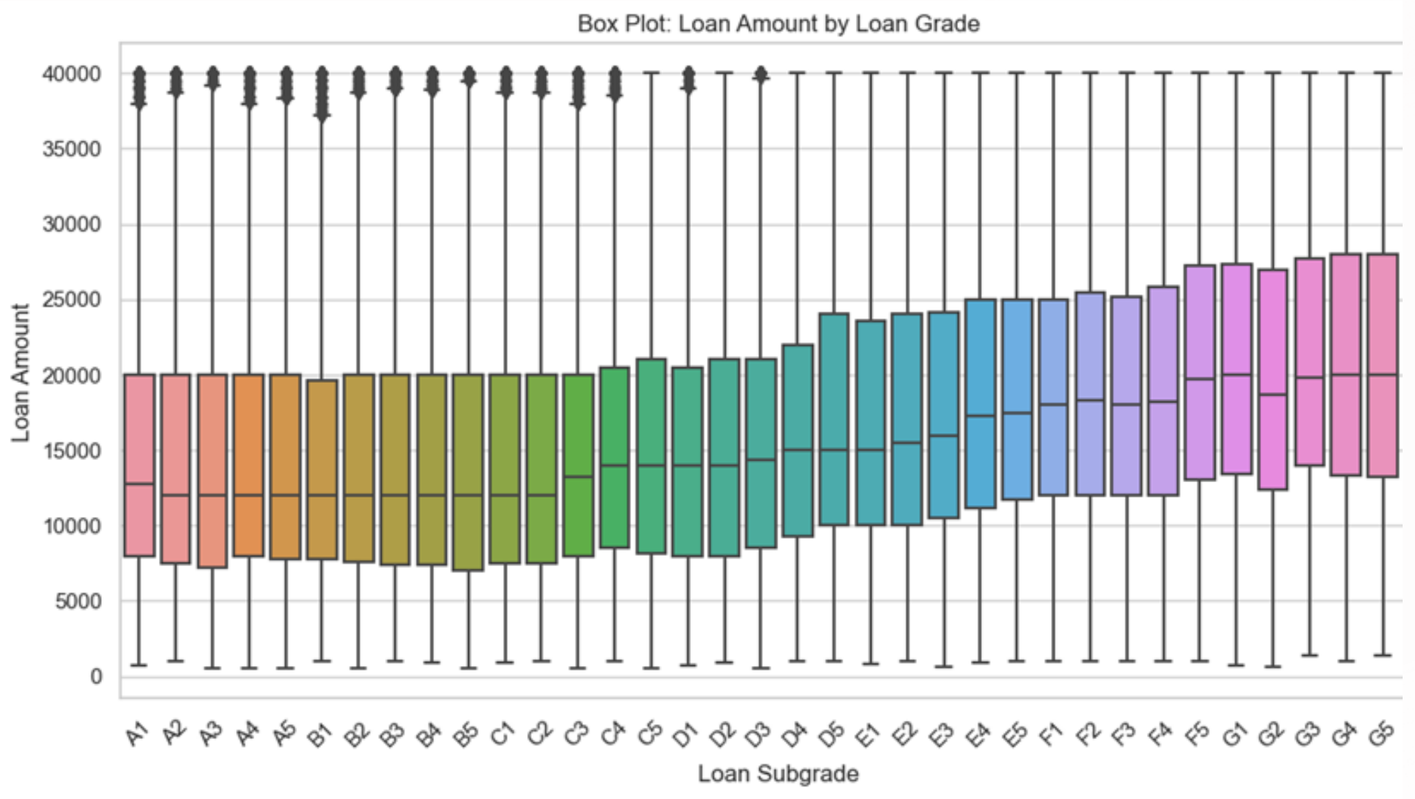
# Loan Status Vs. Loan Grade



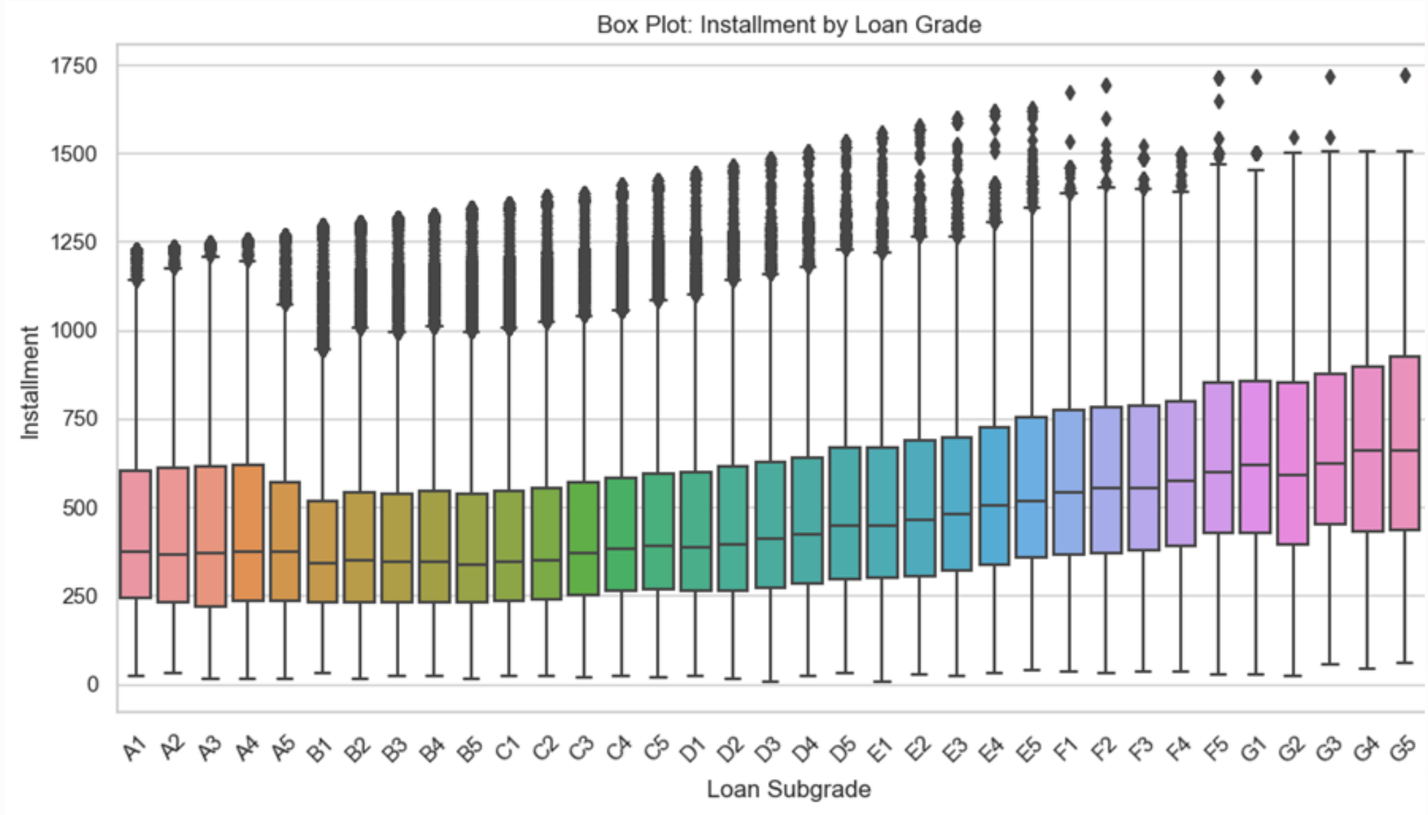
# Interest Rate Vs. Loan Grade



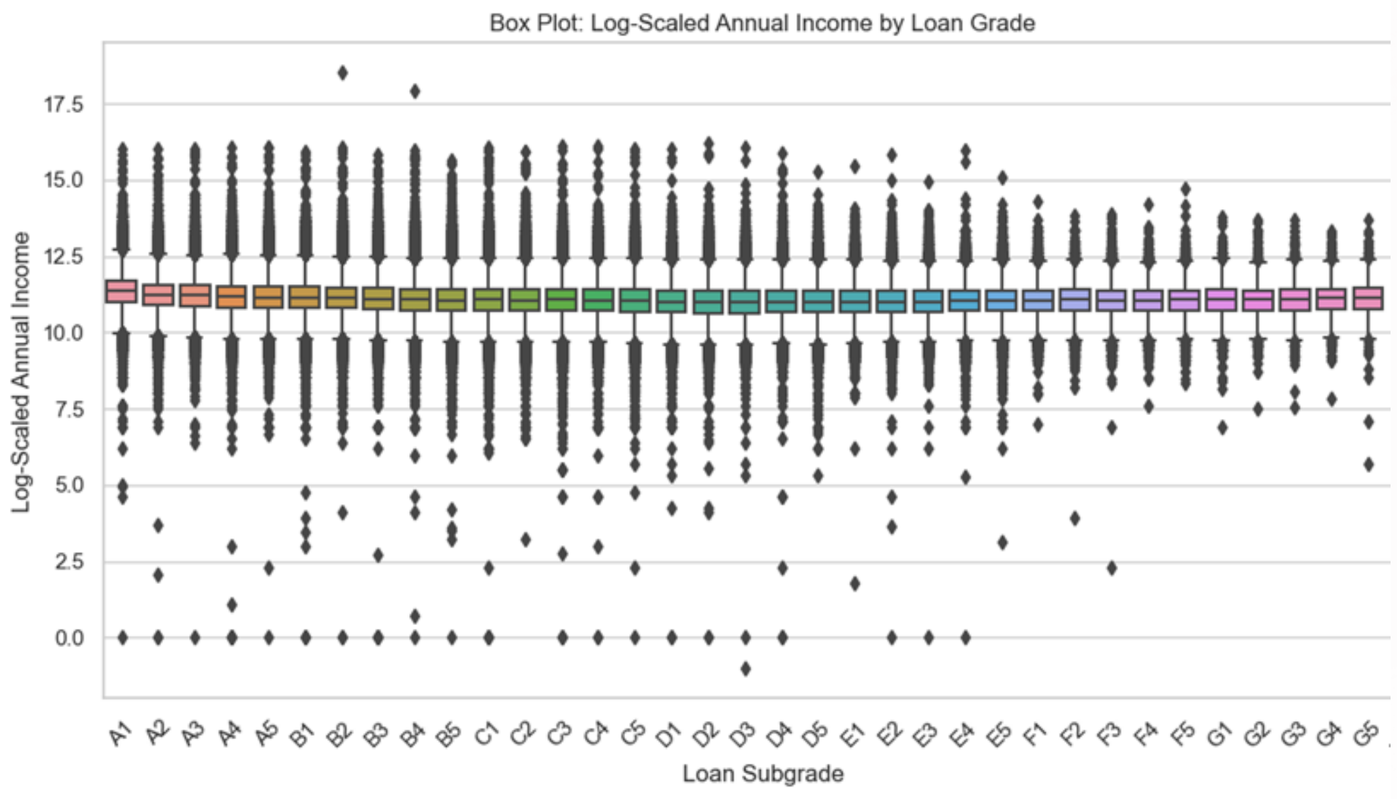
# Loan Amount Vs. Loan Grade



# Loan Installment Vs. Loan Grade



# Annual Income Vs. Loan Grade



# RESULTS & ANALYSIS

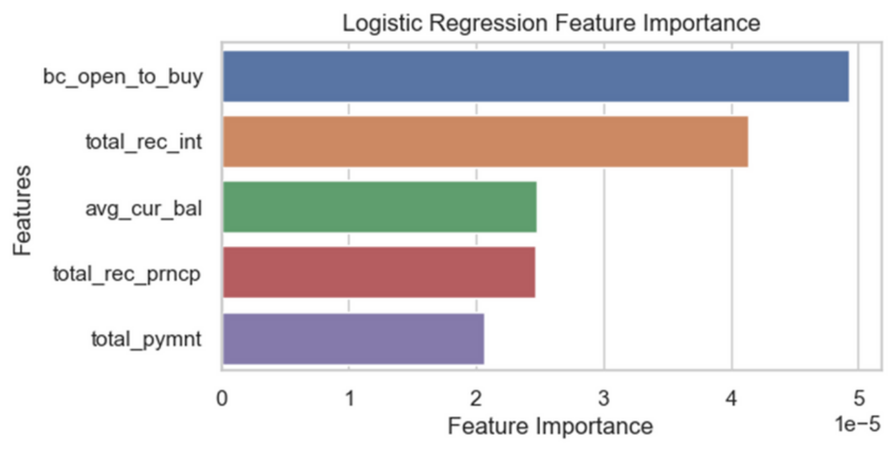
# Model Performance Comparison

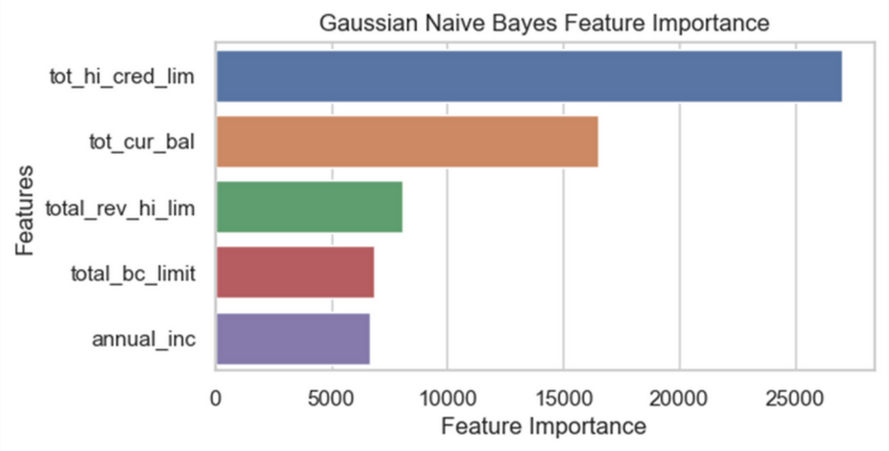
| Algorithm | Testing Accuracy | Weighted Average | | |
|---|---|---|---|---|
| | | Precision | Recall | F1-Score |
| Logistic Regression | 8.86% | 7% | 9% | 6% |
| Gaussian Naïve Bayes | 9.74% | 9% | 10% | 7% |
| Decision Tree | 99.62% | 100% | 100% | 100% |
| Random Forest | 82.15% | 81% | 82% | 82% |
| Bootstrap Aggregating | 99.68% | 100% | 100% | 100% |
| Adaptive Boosting | 17.64% | 13% | 18% | 8% |
| Light Gradient Boosting | 6.28% | 14% | 6% | 3% |
| Extreme Gradient Boosting | 99.71% | 99% | 100% | 99% |

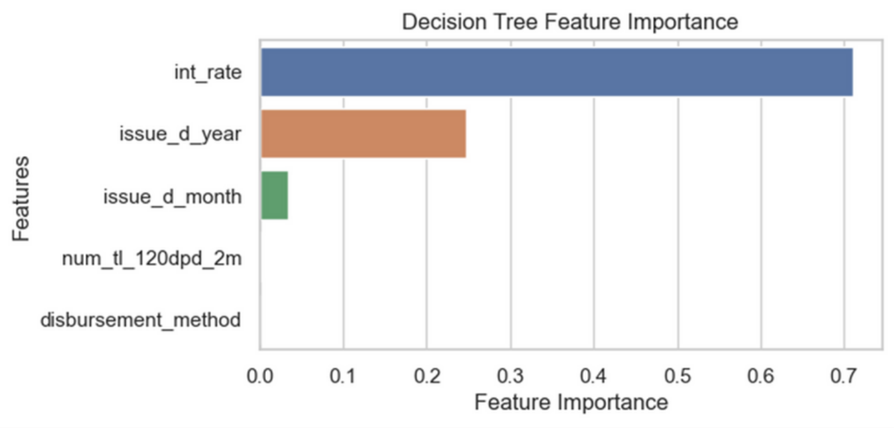
# Time and Cost Efficiency Comparison

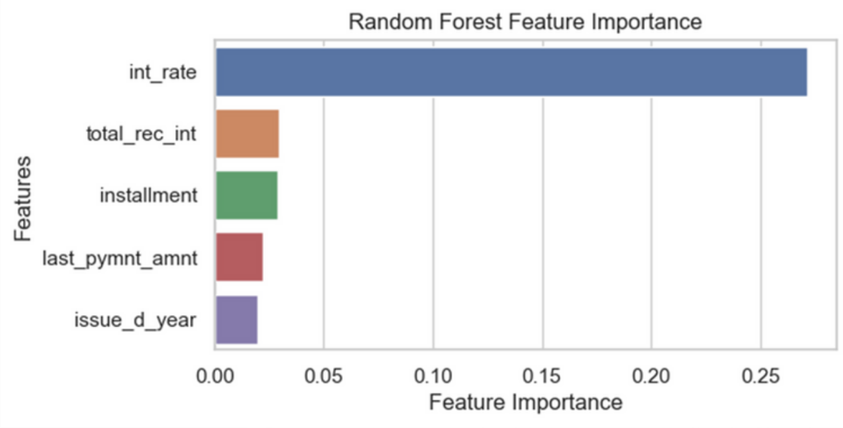
| Algorithms | Training Time |
|---|---|
| Logistic Regession | 3 minutes 42.2 seconds |
| Gaussian Naïve Bayes | 18.9 seconds |
| Decision Tree | 1 minute 32 seconds |
| Random Forest | 34 minutes 46.2 seconds |
| Bootstrap Aggregating | 10 minutes 9.7 seconds |
| Adaptive Boosting | 12 minutes 23.7 seconds |
| Light Gradient Boosting | 2 minutes 53.4 seconds |
| Extreme Gradient Boosting | 71 minutes 25.5 seconds |

# Algorithms Feature Importances

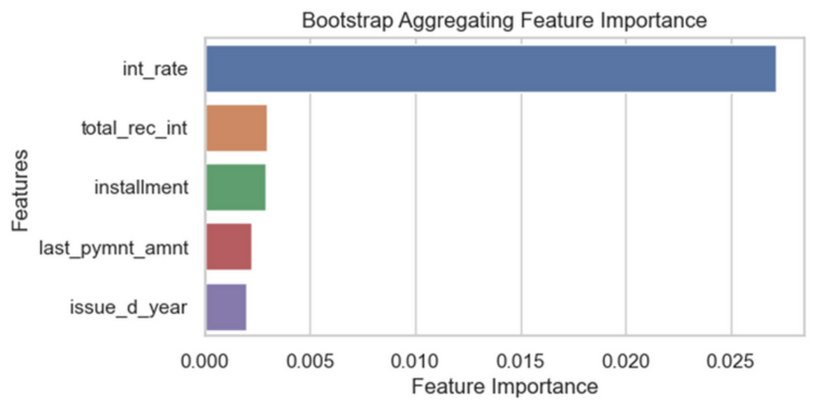


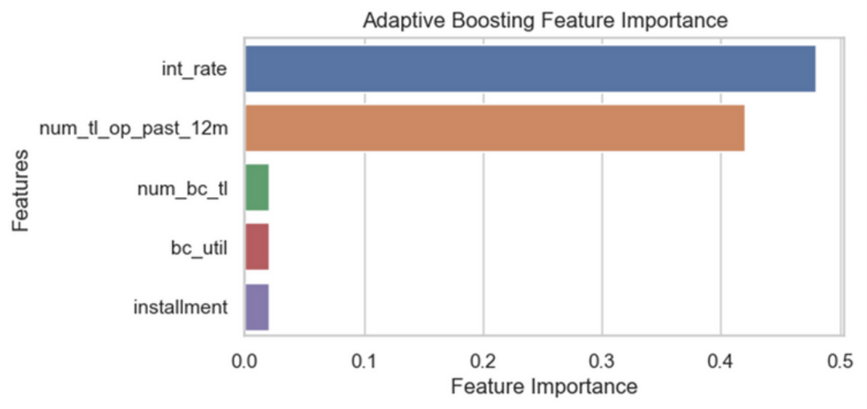



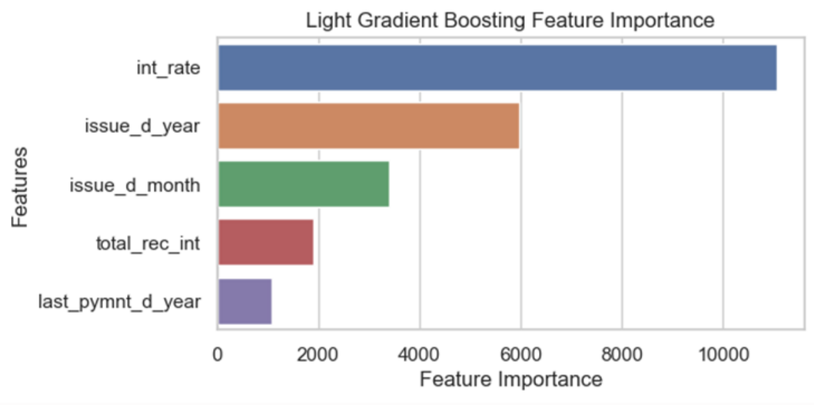



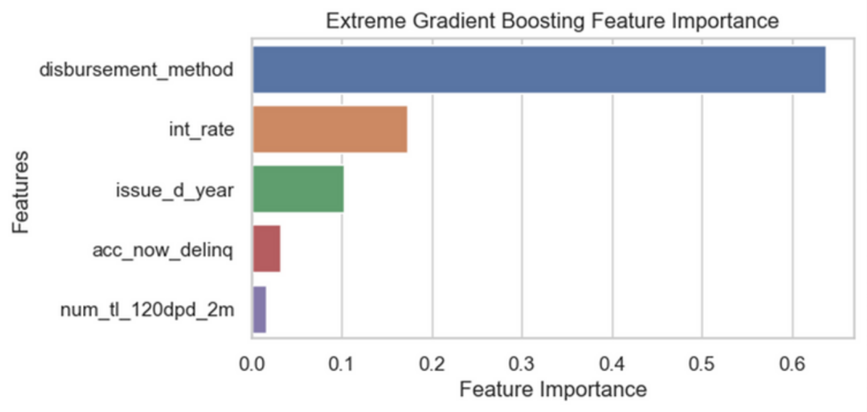



# Algorithms Feature Importances



# CONCLUSION

# Conclusion & Discussion

01

Extreme Gradient Boosting, Decision Tree & Bootstrap Aggregating are the best performing models, with Decision Tree with the most time efficient

02

The other models does not perform good enough to be considered in the real-world application

03

The comparison is not only about the performance of the models, but also the cost that a company has to spend for the model deployment

04

Utilizing technology (machine learning algorithms) may increase a company's productivity (much faster classification & reduce human involvement)

# Implications

- Improve decision making processes
- Enabling effective customer segmentation
- Customize loan offerings based on specific profiles
- Optimize resources allocation

A P U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

# THANK YOU


**Terrence Josiah**

+60173153527

terrence.jo.tan@gmail.com

TP058242

APD3F2211ACS