Capstone Presentation

Bank Loan Default

GREAT LAKES
EXECUTIVE LEARNING

Team BLD-A

Balajisriram Venkateshkumar, Bina Rajput, Narendra Kr. Gupta, Rakendu Sharma, Shailajha Deepak

Agenda

- Team Introduction
- Business Problem Understanding
- Feature Selection and Engineering
- Discussion on Modelling Approach and choice of Final Model
- Insights and Recommendations

Team Introduction

Bina Rajput

Shailajha Deepak

Narendra Kr. Gupta

Balajisriram Venkateshkumar

Rakendu Sharma

Business Problem Understanding

Loan default is a major risk faced by banks and financial institutions since it impacts profitability. Our goal is to help banks and financial institutions minimize defaults and improve bottom line:

Rejecting a customer with a good credit profile assuming they will default resulting in loss of business to the bank.

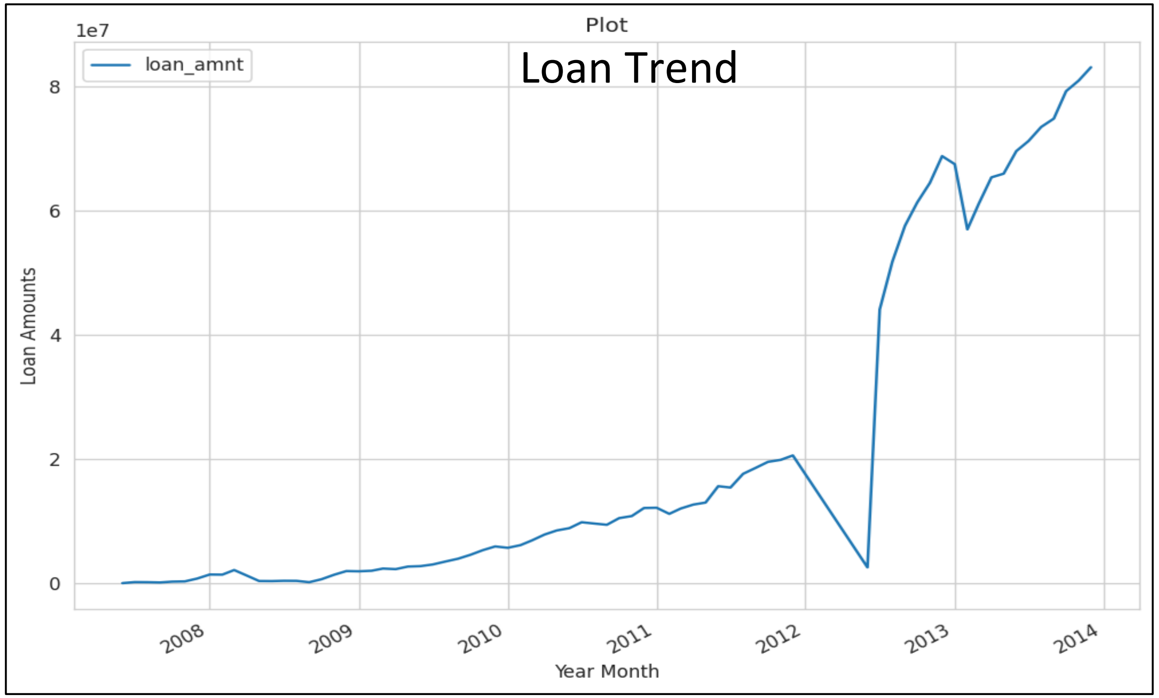
Approving a customer with a bad credit profile without realizing the customer may default, which may result in high losses if the customer defaults.

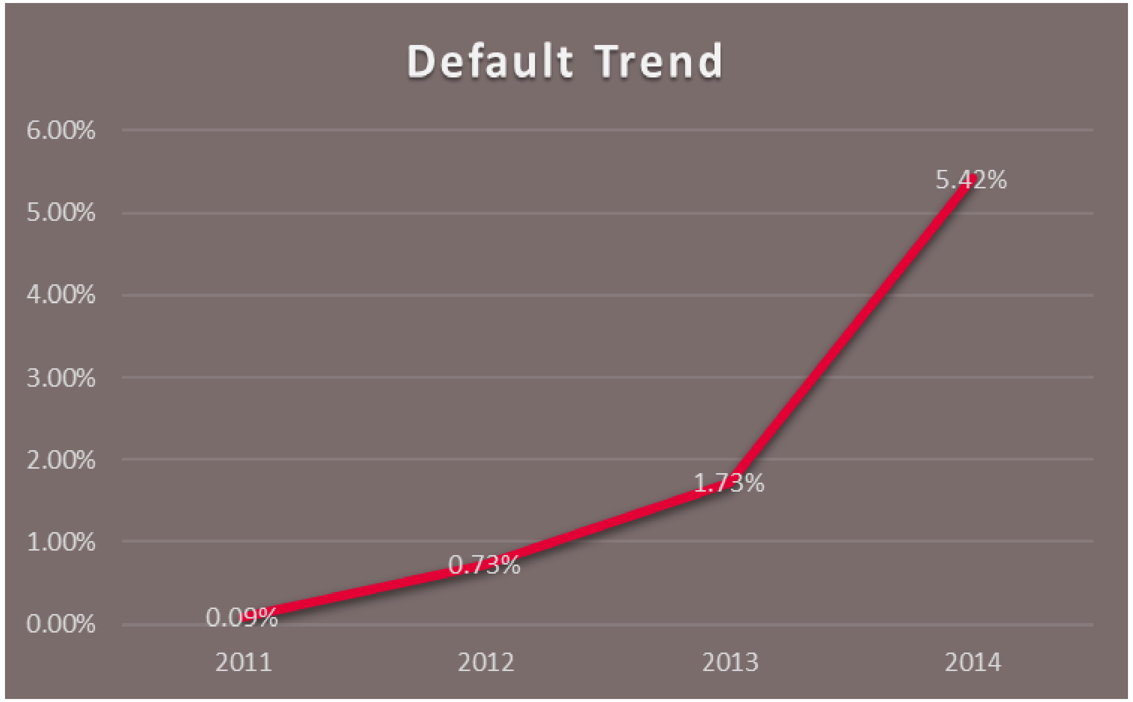
Important to use the capabilities of Data Analytics and build robust machine learning prediction models to support growth and profitability of banks.

Classification Algorithms appropriate for the Business Prediction Model.

Business Problem Understanding

Decisions based on balancing Risks and Rewards





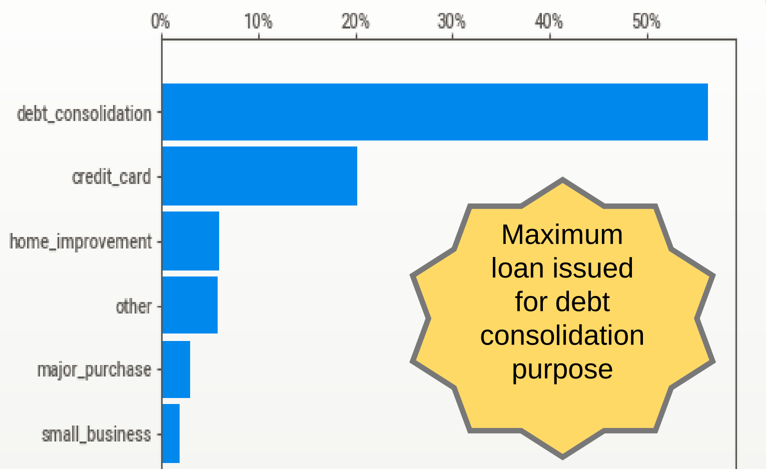
Demand of loan increased significantly in last two years. Need for proper risk evaluation is essential.

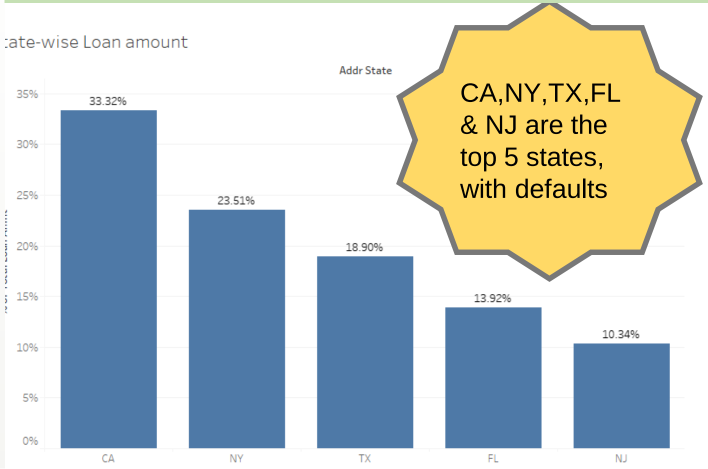
In the same period, the percentage of loan defaulters also increased considerably.

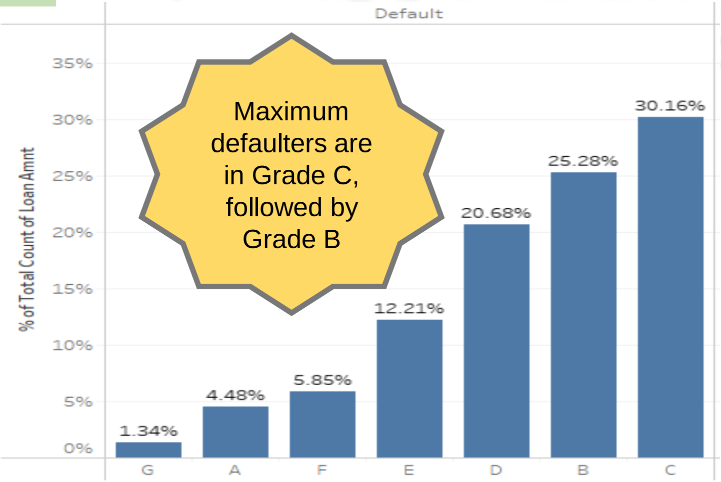
Projected to grow higher if not mitigated strategically with a default prediction model.

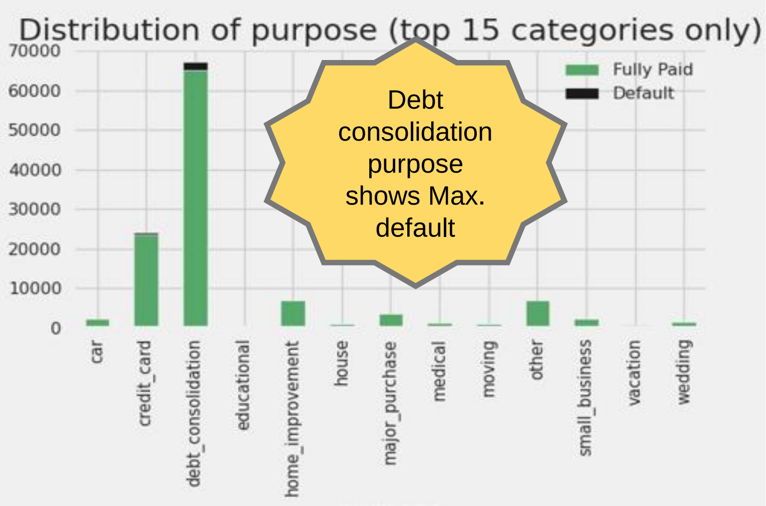
Feature Selection and Engineering

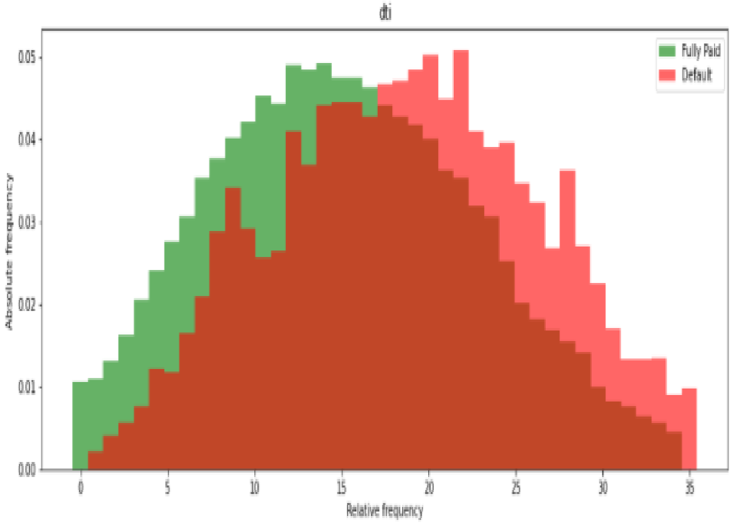
Exploratory Data Analysis

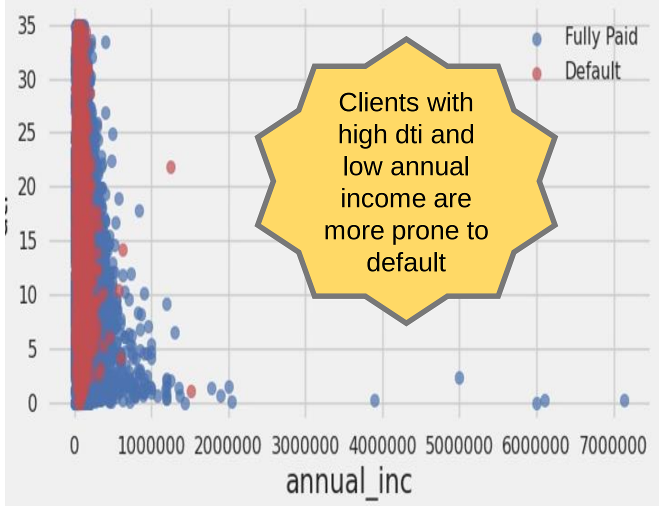












Feature Selection and Engineering

Feature Selection

Based on the Target Variable **Loan_Status** (Fully Paid / Default) - **Supervised Method**

Filtered the provided data based on the **Pre-facto & Post Facto- variables**

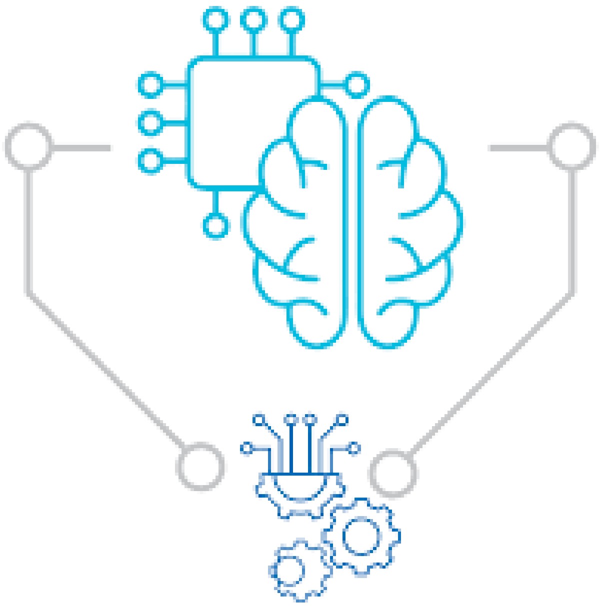
Variables available in Loan application form are considered here (*Out of 41 Columns only 17 columns are used as Input variable for model prediction)*

- √ Loan_amnt
- √ funded_amnt_inv
- √ term
- √ int_rate
- √ Installment
- √ Grade
- √ Emp_length
- √ Home_ownership
- √ annual_inc
- √ Purpose
- √ dti Open_acc
- √ revol_bal
- √ revol_util
- √ Total_acc
- √ addr_state
- √ Loan_status (TV)

Feature Selection and Engineering

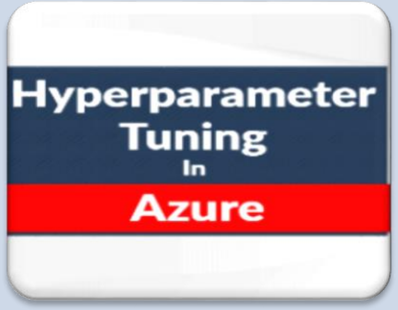
Feature Engineering

- **Data wrangling**
  - Cleaning & mapping the data to fit for modelling
- **Data Imputation**
  - Numerical Imputation: Replacing the missing value
- **Balancing data**:
  - Multi Collinearity
  - SMOTE
  - Skewness or Degree of Distortion to normalize data



The framing of the problem : **Classification Problem**

Discussion on Modelling Approach



Run model on batches

Model run in a iterative process to check on **Overfitting Or Under fitting** models

Pre facto/ Post facto

Variables during loan application

Variables after loan disbursement

Data Splitting

The data split into Training and Test Ratio of 70:30 with stratified split to handle imbalance

SMOTE

High Imbalance in Dependent variable data (97% vs 3%) - SMOTE function

Hyper parameter

Optimize performance on the data in a reasonable amount of time:

Random Sweep

Random Grid

Evaluation criteria

F1 - Score

AUC

Confusion Matrix

Choice of Final Model

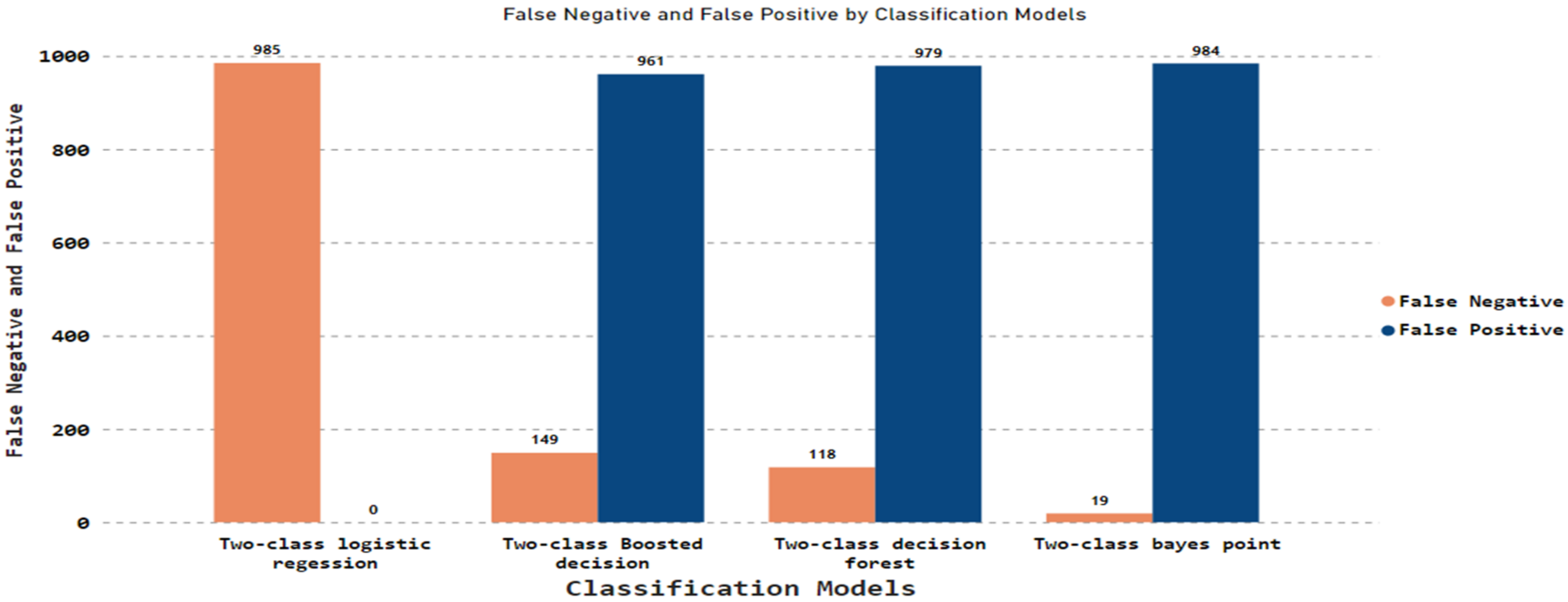
Dependable Variable – Categorical – Fully Paid and Default

Two-class classification model vs Multi-class classification model

Two-class Classification Models

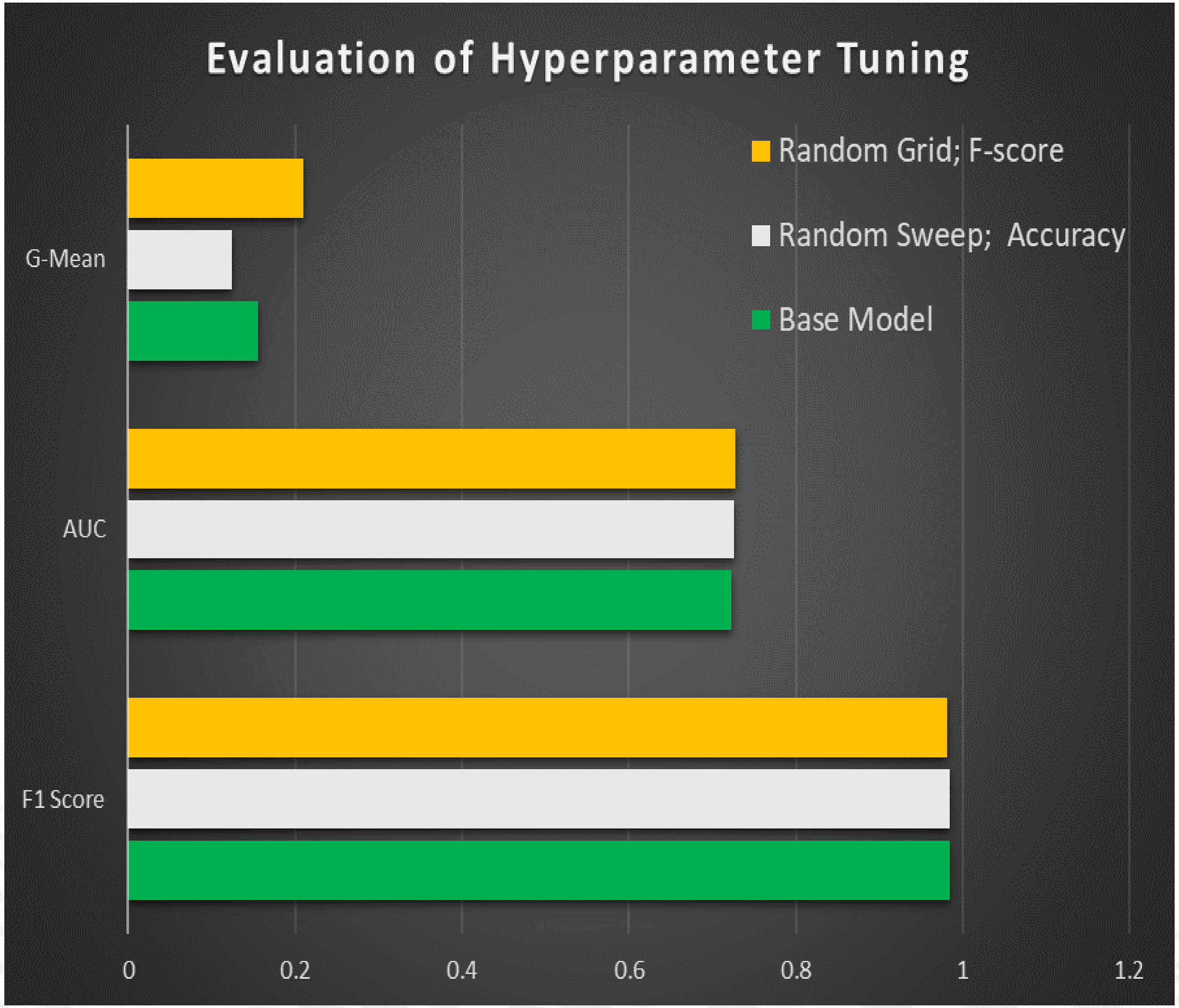
- **Boosted Decision**
- **Decision Forest**
- **Bayes Point**
- **Logistic Regression**

Choice of Final Model



Classification Models	Confusion Matrix			
	True Positive	False Positive	True negative	False Negative
Two-class Boosted decision	34610	961	24	149
Two-class decision forest	34641	979	6	118
Two-class logistic regession	34759	0	0	985
Two-class bayes point	34740	984	1	19

Hyperparameter Tuning



Hyperparameter tuning for the model Two-class boosted decision tree.

No significant difference in the selected criteria after tuning.

Two Class Boosted Decision Tree itself is a Tuned Ensemble model in itself

Insights

- Model prediction will help identify defaulters and reduce losses
- The data pertains to consumer loans. Outliers are detected in the data
- Proper risk benefit evaluation for higher default categories
- High defaulters in Debt-consolidation category, state of California, dti ranging between 15 to 20
- Customers with property on mortgage, multiple credit lines, high instalments also have higher defaults
- Some categories of loans are perceived to be risk free

Recommendations

- Banks to explore growth opportunities through proper risk benefit evaluation
- High collateral or guarantee and higher interest can be undertaken for high risk borrowers to mitigate the losses
- Regionally tailored risk assessment and policies could potentially achieve more accurate default
- Good quality data to build a more robust prediction model
- Data related to customer such as age, gender, number of dependants, heuristics to help build a better model

Thank you

vivid analytica

Analytics at its finest