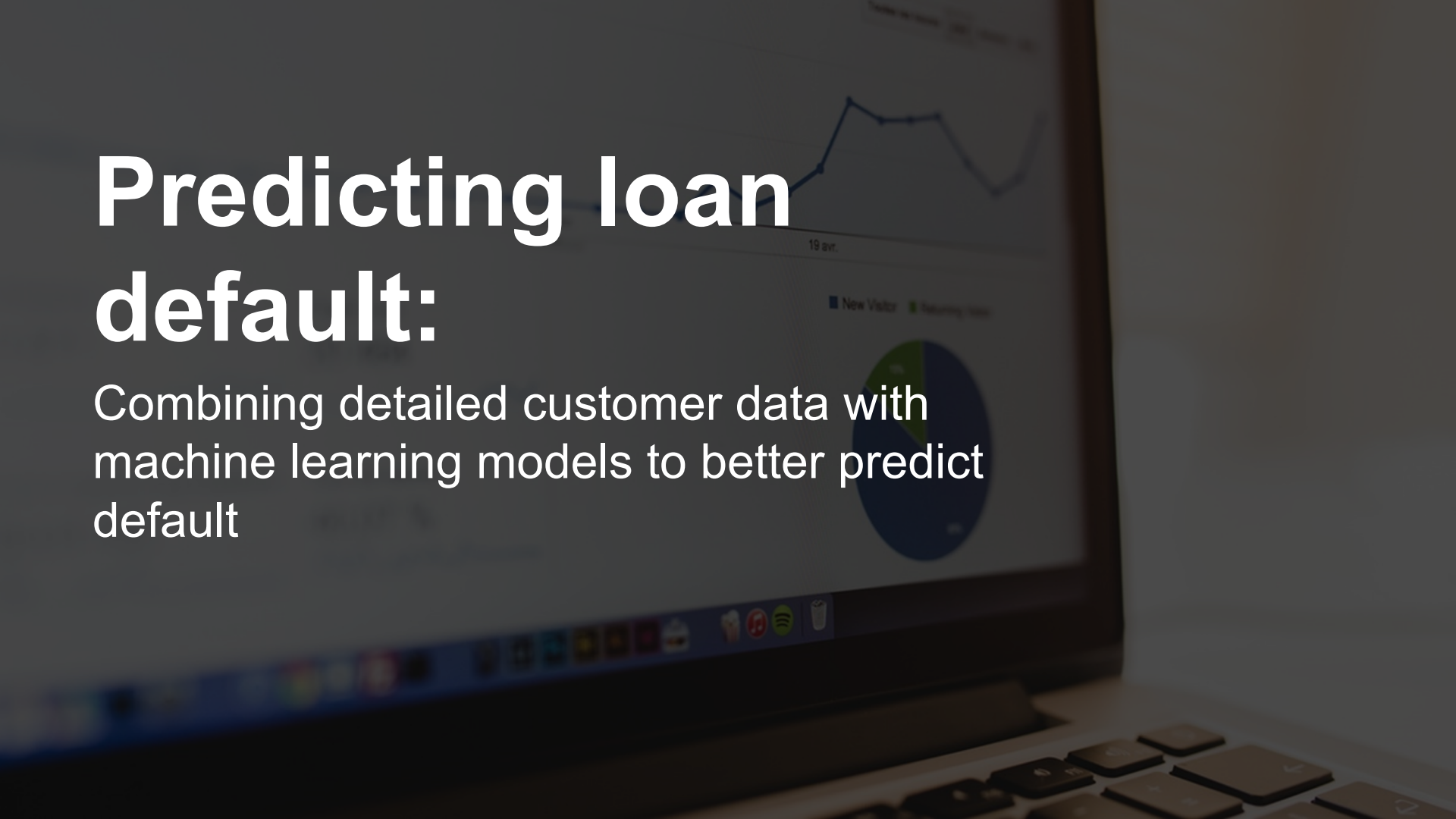


# Predicting loan default:

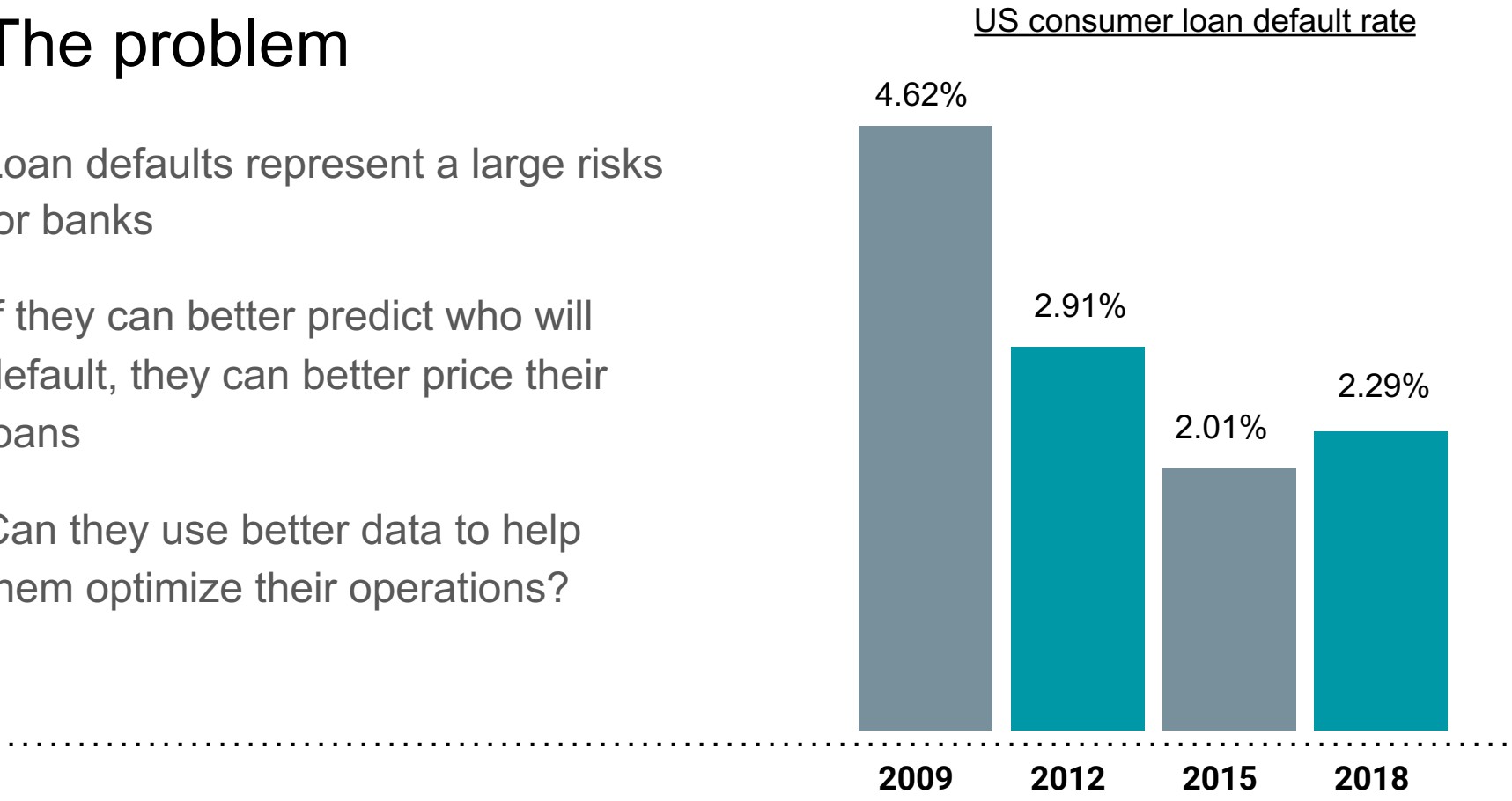
Combining detailed customer data with machine learning models to better predict default

# The problem

Loan defaults represent a large risks for banks

If they can better predict who will default, they can better price their loans

Can they use better data to help them optimize their operations?



# Data Sources and Tools

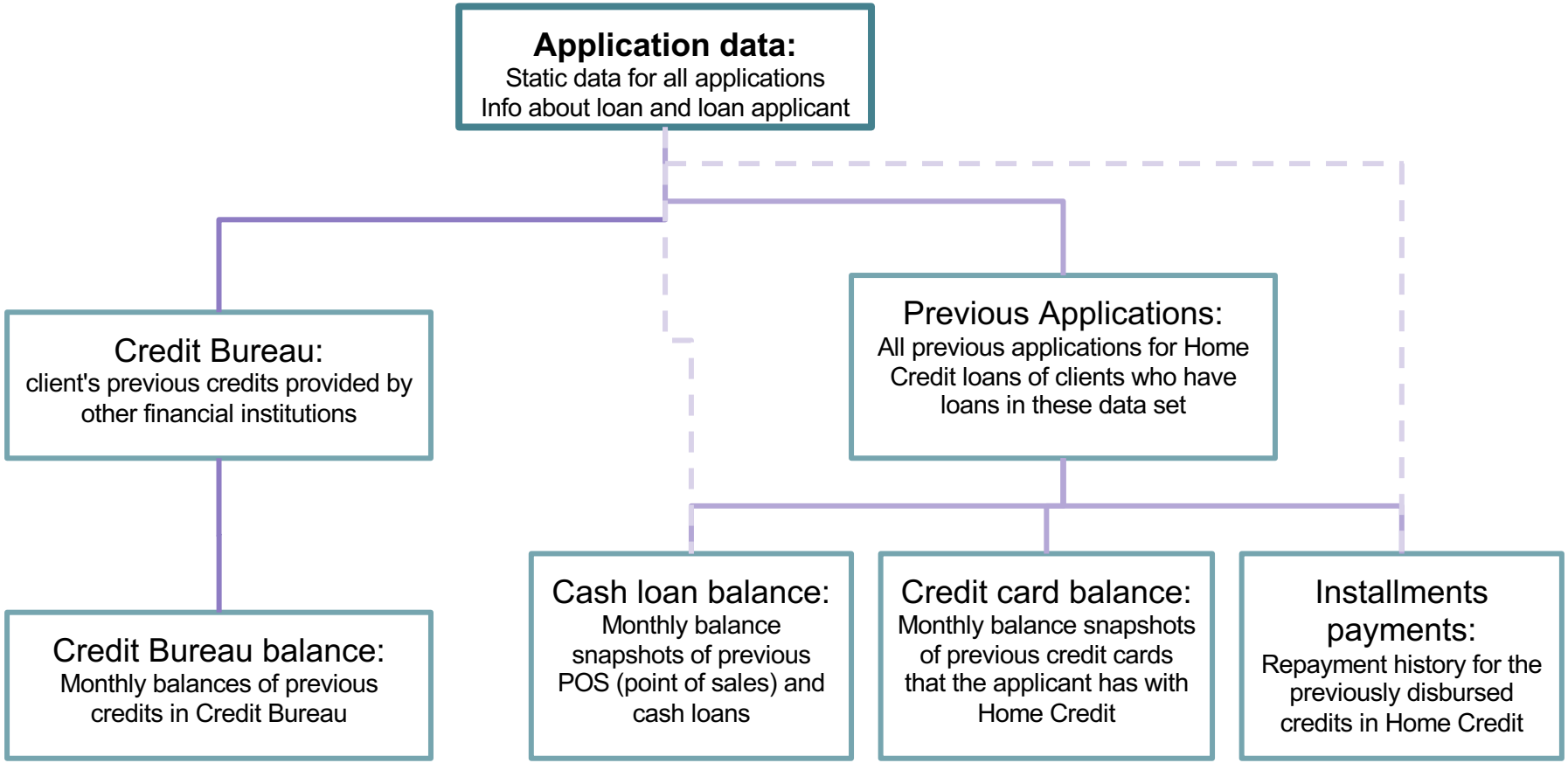
Loan Data provided by Home Credit Bank.

Competition - the winning prize of $70K!

Can I get close to the winning score?!





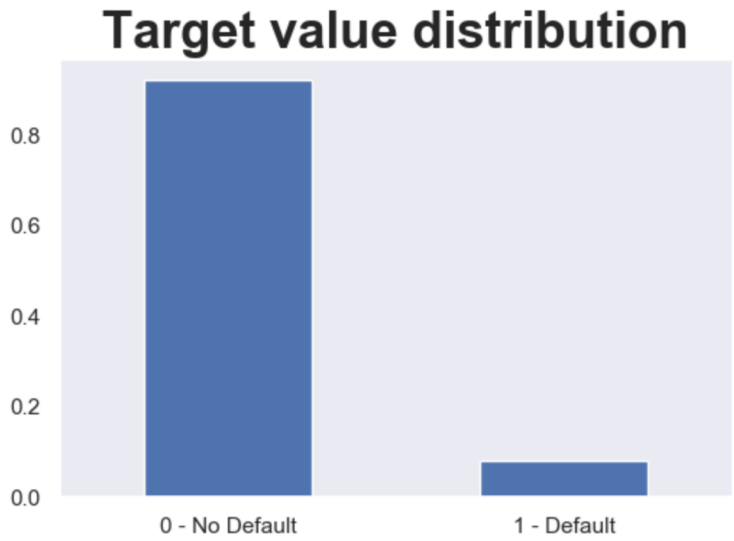

Application data:
Static data for all applications
Info about loan and loan applicant
Credit Bureau:
client's previous credits provided by other financial institutions
Credit Bureau balance:
Monthly balances of previous credits in Credit Bureau
Previous Applications:
All previous applications for Home Credit loans of clients who have loans in these data set
Cash loan balance:
Monthly balance snapshots of previous POS (point of sales) and cash loans
Credit card balance:
Monthly balance snapshots of previous credit cards that the applicant has with Home Credit
Installments payments:
Repayment history for the previously disbursed credits in Home Credit

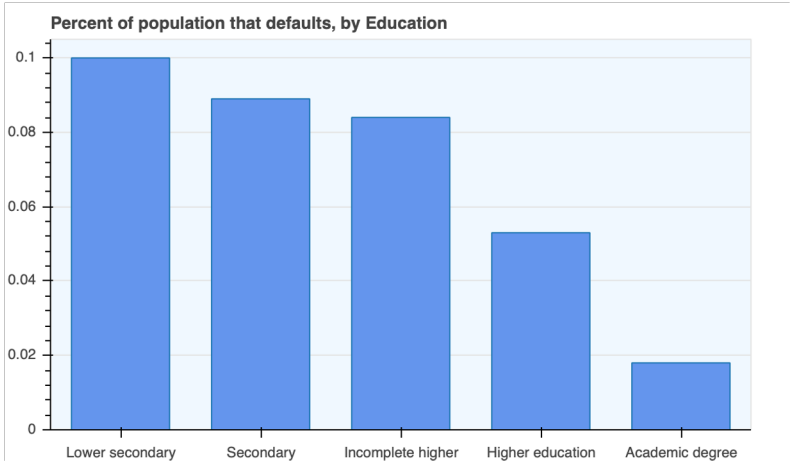
**Some of the selected features include:**

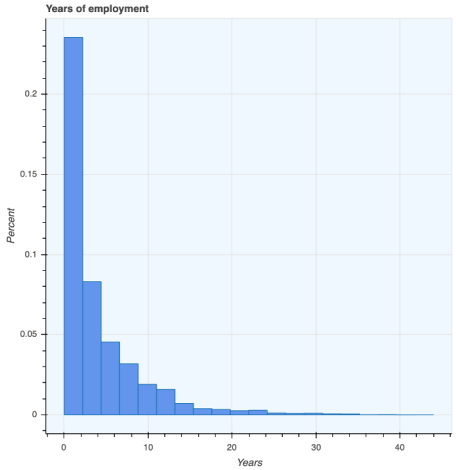
- Client’s monthly income;
- Monthly payments;
- Whether client owns a real estate;
- Whether client could be reached via phone;
- Education level;
- Payment history, if any.

**Some created features include:**

- Income to Debt ratio;
- Average number of applications that client had previously;
- Rejection rate if client applied for the loan before;
- Average number of days that client had past due



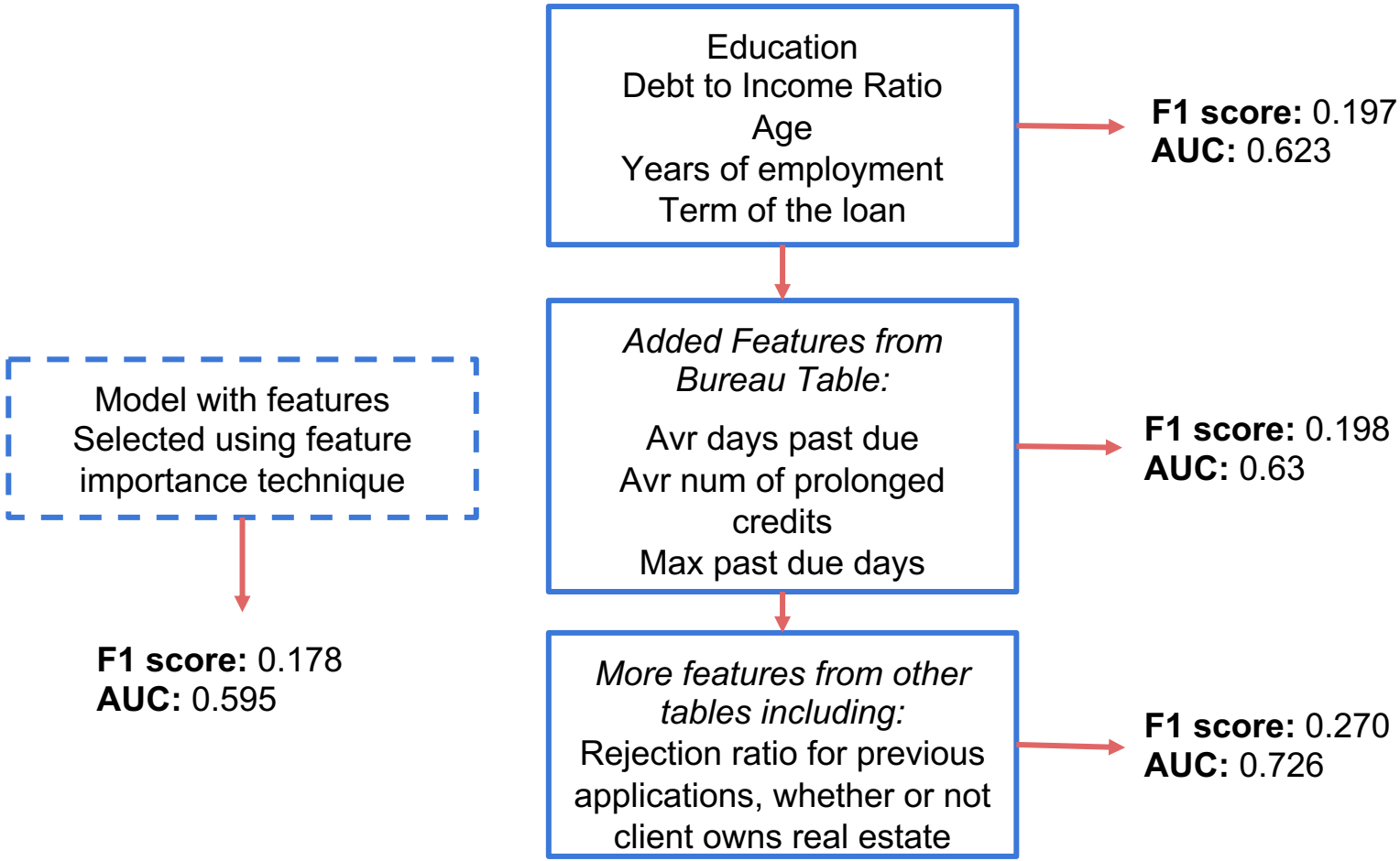

**Data is imbalanced**

Metrics for analysis:

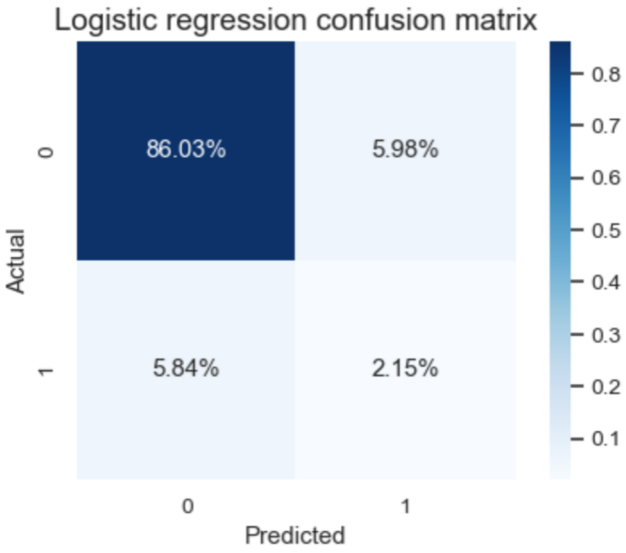
**F1 score** and **AUC**

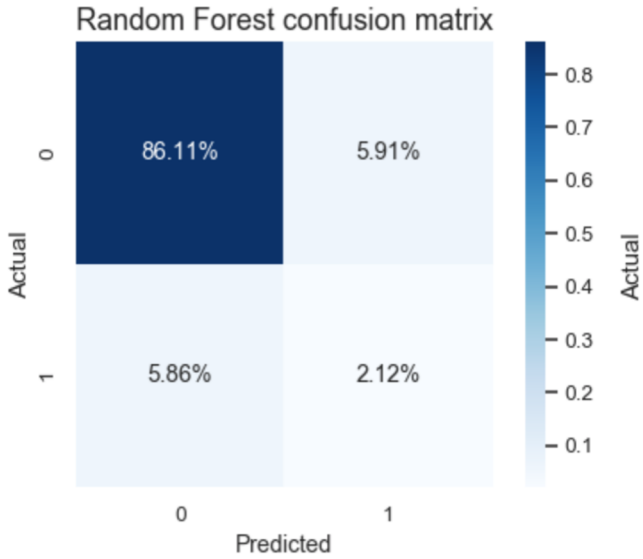
Will be using soft classification

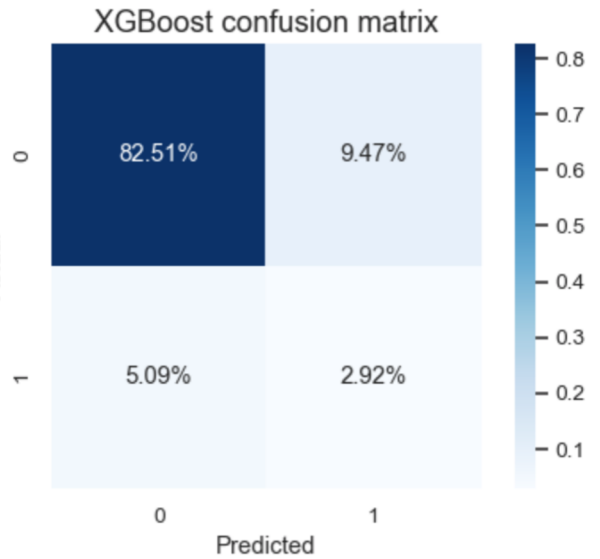
# Base model: Logistic regression

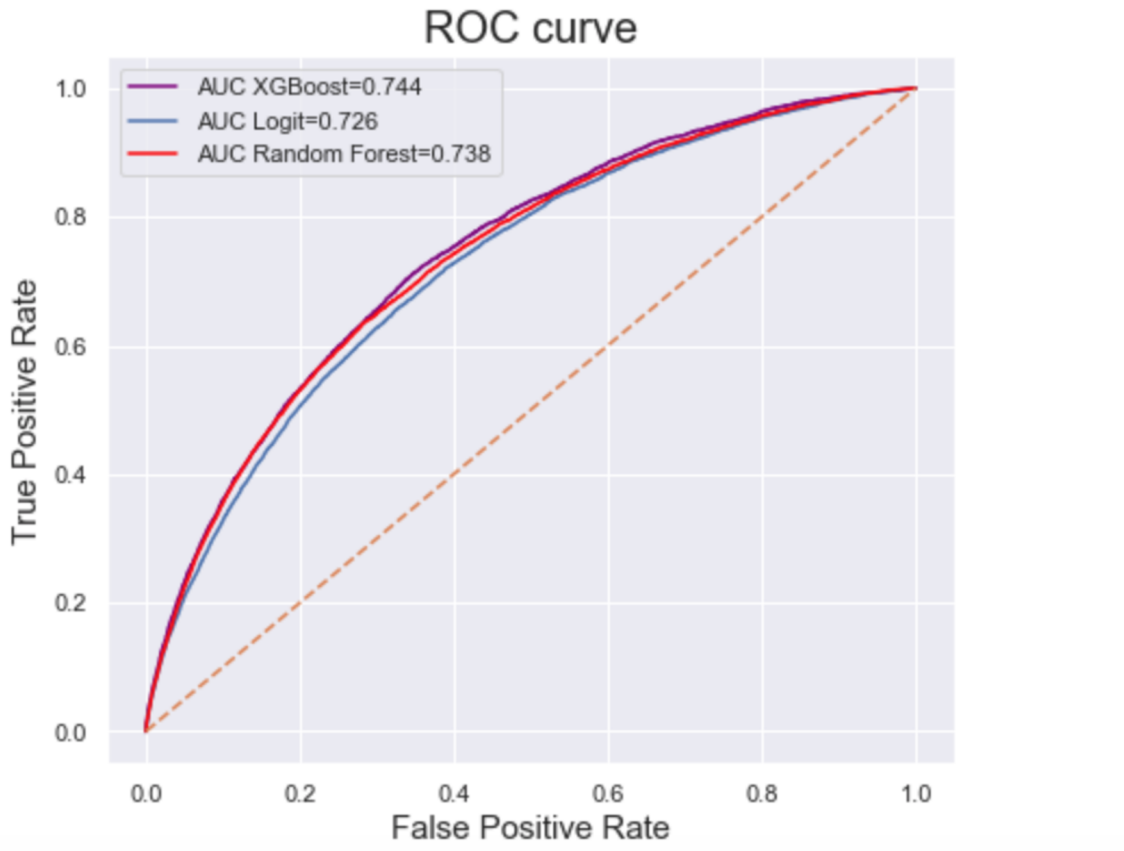


| | F1 Score | AUC |
|---|---|---|
| **Logistic regression** | 0.270 | 0.726 |
| **Random Forest** | 0.286 | 0.737 |
| **XGBoost** | 0.286 | 0.74 |




ROC curve
AUC XGBoost=0.744
AUC Logit=0.726
AUC Random Forest=0.738
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

# Results and improvements:

1. The winning score for the competition was 0.8.  My model had a score of 0.74, which is a good start but also leaves room for improvement

2. To improve my model I would go again through my data and use more feature engineering.