

# Home Credit Default Risk Detection

*An Application of Big Data &  
Artificial Intelligence*

**Group Name:** Map Reducers

**Group Members:** Sahitya Angara, Akshay Havalgi, Mit Patel, Shashank Manu Rao, Vivek Sivalingam, Nicholas Summers



# Hello!

**We are the Map Reducers.**

A team of analysts inspired and driven by the infinite potential of Data Science.

Sahitya A



Nick S



Akshay H



Vivek S



Mit P



Shashank R



# What is Home Credit Default Risk?

- Home Credit risk is the probability of a client not being able pay back a housing loan on time
- This is a problem of high importance in the housing sector as well as the financial sector

## ○ Why is credit risk important?

- Credit defaults affect the economy (eg: 2008 recession)
- Prevent lenders from lending to certain borrowers
- Prevent borrowers from borrowing too much

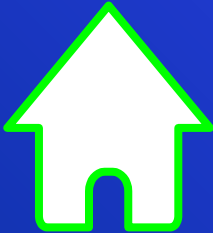
# 16.1 Trillion USD



Whoa! That's a lot of money, is this good?

# The Impact of Home Credit Risk

How does this affect everyone?



# Damage to Various Stakeholders



Borrowers can lose their assets when they default on a loan. They can also declare bankruptcy.

Borrowers can have their wages garnished.

Lending institutions can lose their financial asset.

Lending institutions can become insolvent.



# Our Approach

## Our Data

- **Where:** Kaggle
- **What:** Data of Loan Applicants
- **Why:** The large feature set; It was a BIG DATA set - Good potential for EDA, modeling, and extracting practical insights



## Data ETL

- Kaggle → Google Colab
- Joined datasets based on FKs
- Performed EDA
- Identified and eliminated outliers
- Feature Engineering
- Loaded in G Drive



## Modeling

- Model selection
- Regressions
- Clustering
- Neural Nets
- Boosting



# Our process is simple

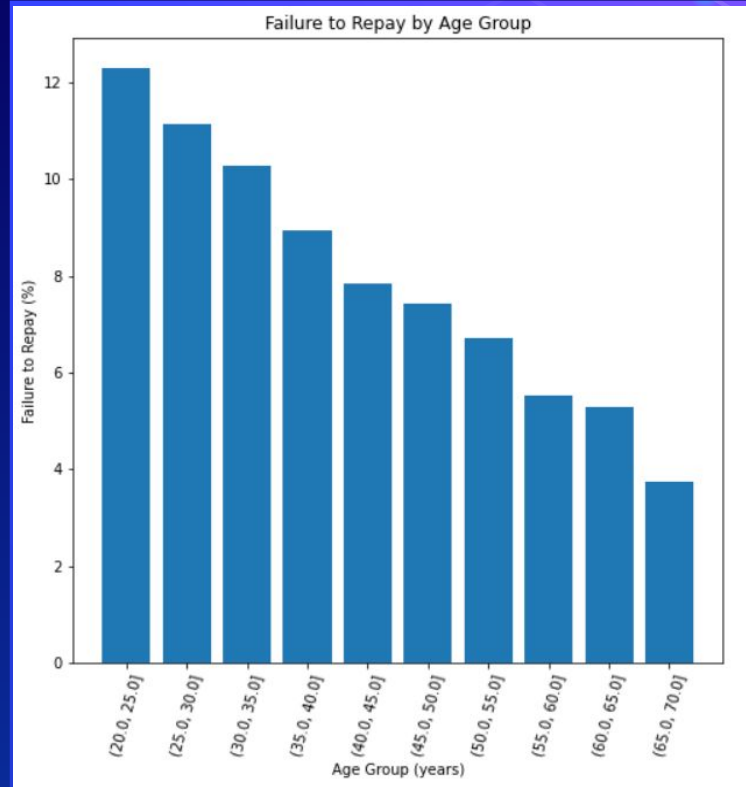




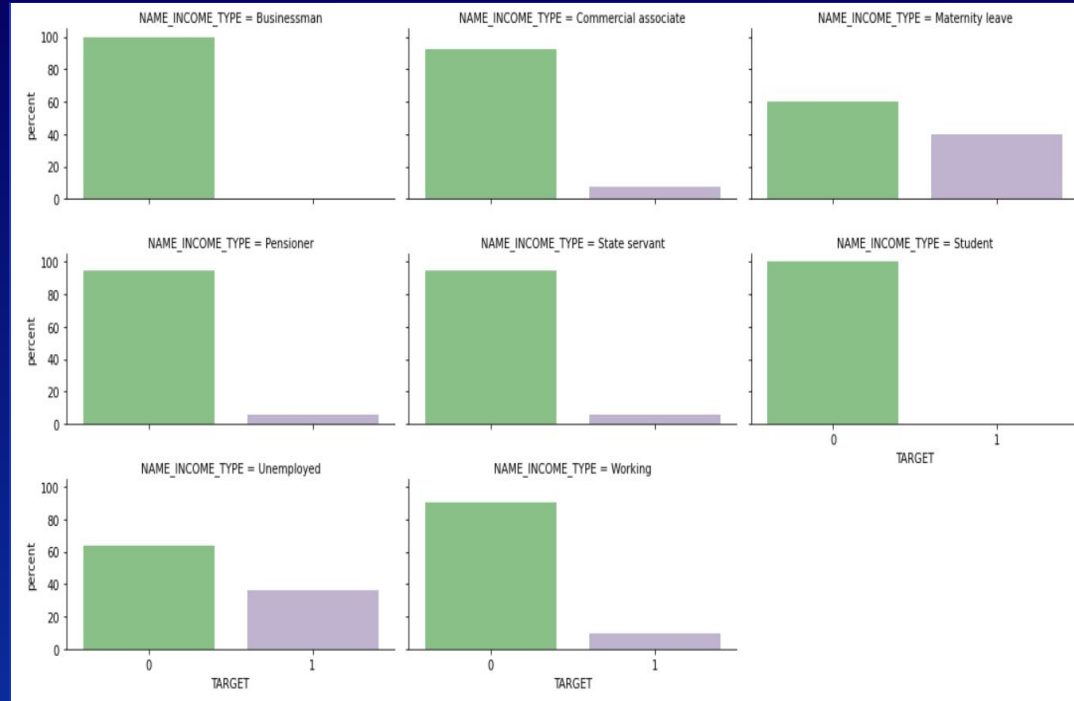
# Exploratory Data Analysis



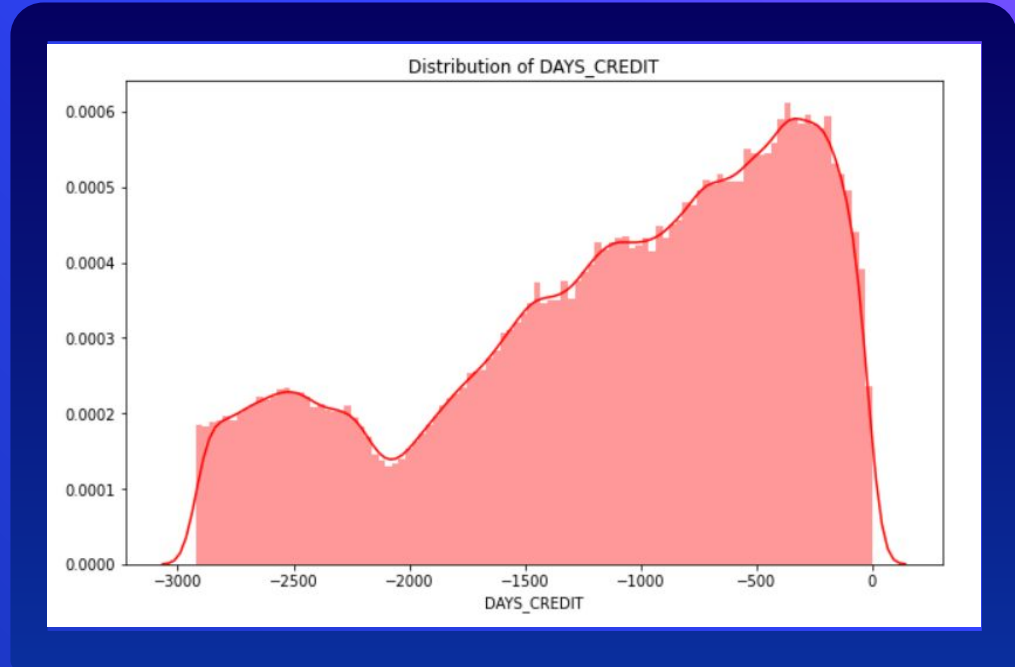
# Effect of age on repayment



# Influence of Income Type on Credit Risk



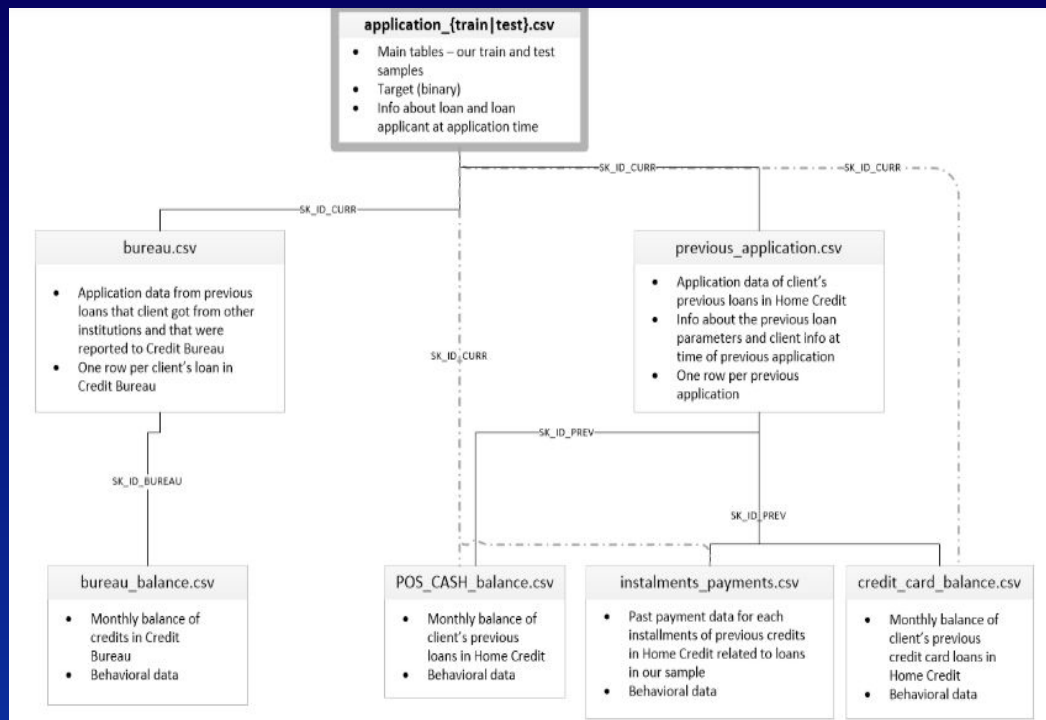
# Distribution of Duration between Credit Loan Application and Home Credit Application



# Extract Transform Load



# Data Structure





# Manual Feature Engineering

```
#Customized features
app['LOAN_RATE'] = app['AMT_ANNUITY'] / app['AMT_CREDIT']
app['CREDIT_INCOME_RATIO'] = app['AMT_CREDIT'] / app['AMT_INCOME_TOTAL']
app['EMPLOYED_BIRTH_RATIO'] = app['DAYS_EMPLOYED'] / app['DAYS_BIRTH']
app['EXT_SOURCE_SUM'] = app[['EXT_SOURCE_1', 'EXT_SOURCE_2', 'EXT_SOURCE_3']].sum(axis = 1)
app['EXT_SOURCE_MEAN'] = app[['EXT_SOURCE_1', 'EXT_SOURCE_2', 'EXT_SOURCE_3']].mean(axis = 1)
app['AMT_REQ_SUM'] = app[[x for x in app.columns if 'AMT_REQ_' in x]].sum(axis = 1)

bureau['LOAN_RATE'] = bureau['AMT_ANNUITY'] / bureau['AMT_CREDIT_SUM']

bureau_balance['PAST_DUE'] = bureau_balance['STATUS'].isin(['1', '2', '3', '4', '5'])
bureau_balance['ON_TIME'] = bureau_balance['STATUS'] == '0'

previous['LOAN_RATE'] = previous['AMT_ANNUITY'] / previous['AMT_CREDIT']
previous['AMT_DIFFERENCE'] = previous['AMT_CREDIT'] - previous['AMT_APPLICATION']

installments['LATE'] = installments['DAYS_ENTRY_PAYMENT'] > installments['DAYS_INSTALLMENT']
installments['LOW_PAYMENT'] = installments['AMT_PAYMENT'] < installments['AMT_INSTALLMENT']

cash['LATE_PAYMENT'] = cash['SK_DPD'] > 0.0
cash['INSTALLMENTS_PAID'] = cash['CNT_INSTALLMENT'] - cash['CNT_INSTALLMENT_FUTURE']


credit['OVER_LIMIT'] = credit['AMT_BALANCE'] > credit['AMT_CREDIT_LIMIT_ACTUAL']
credit['BALANCE_CLEARED'] = credit['AMT_BALANCE'] == 0.0
credit['LOW_PAYMENT'] = credit['AMT_PAYMENT_CURRENT'] < credit['AMT_INST_MIN_REGULARITY']
credit['LATE'] = credit['SK_DPD'] > 0.0
```

# Machine Learning

Using machine learning to detect home credit default risk and derive insights to drive business decisions



# Which to choose ?

Model	K-Means Clustering	Logistic Regression	Random Forest	 XGBoost	Conv. Neural Network
Accuracy	55.7%	70.3%	69.0%	70.5%	68.5%
True Positive Rate	61.3%	69.9%	70.3%	70.1%	71.2%

# Model Selection

## Best Models

(Accuracy; TPR)

- Logistic Regression
- XGBoost

## And the winner is..

- XGBoost
- **Why:** Low chances of overfitting, automatic feature selection

## Key Constraints

- False Negative - **Undesirable**  
(mistaking defaulters for non-defaulters)
- True Positives & Accuracy:  
Most important factors for consideration



“

So what ?!

# Our Findings



The age of the borrower can impact their probability of default on a loan; Young people tend to default at higher rates.



Accurate credit risk assessments can protect the financial well-being of both borrowers and lenders as well as the overall health of the economy.



# Who would benefit from this analysis?

## Financial Services Sector

- The total mortgage debt outstanding in the U.S. amounted to approximately 16.01 trillion U.S. dollars in 2019.
- Defaults on these could devastate a bank and the economy.

## The Government

The U.S Federal Government wants to ensure that banks loan responsibly in order to prevent another economic crisis like we experienced in 2008.

## The Consumer

If financial institutions improve their credit risk assessments, borrowers will only be able to borrow money that they can pay back reasonably. This can prevent consumers from losing their homes due to foreclosure.

# Thanks!

We'll be happy to answer any questions you may have! 👍

