Context and Goals

- **Context**
  - Given problem was about loan defaults. Bank provides home equity line of credit to customers
  - Certain number of customers do not pay back loan or are seriously delinquent. This eats into bank’s profits.
  - If number of loan defaults is reduced through automated process, it would improve bank’s profits and may also help in increasing actual number of loans in long run when bank is confident in identifying potential bad loans.

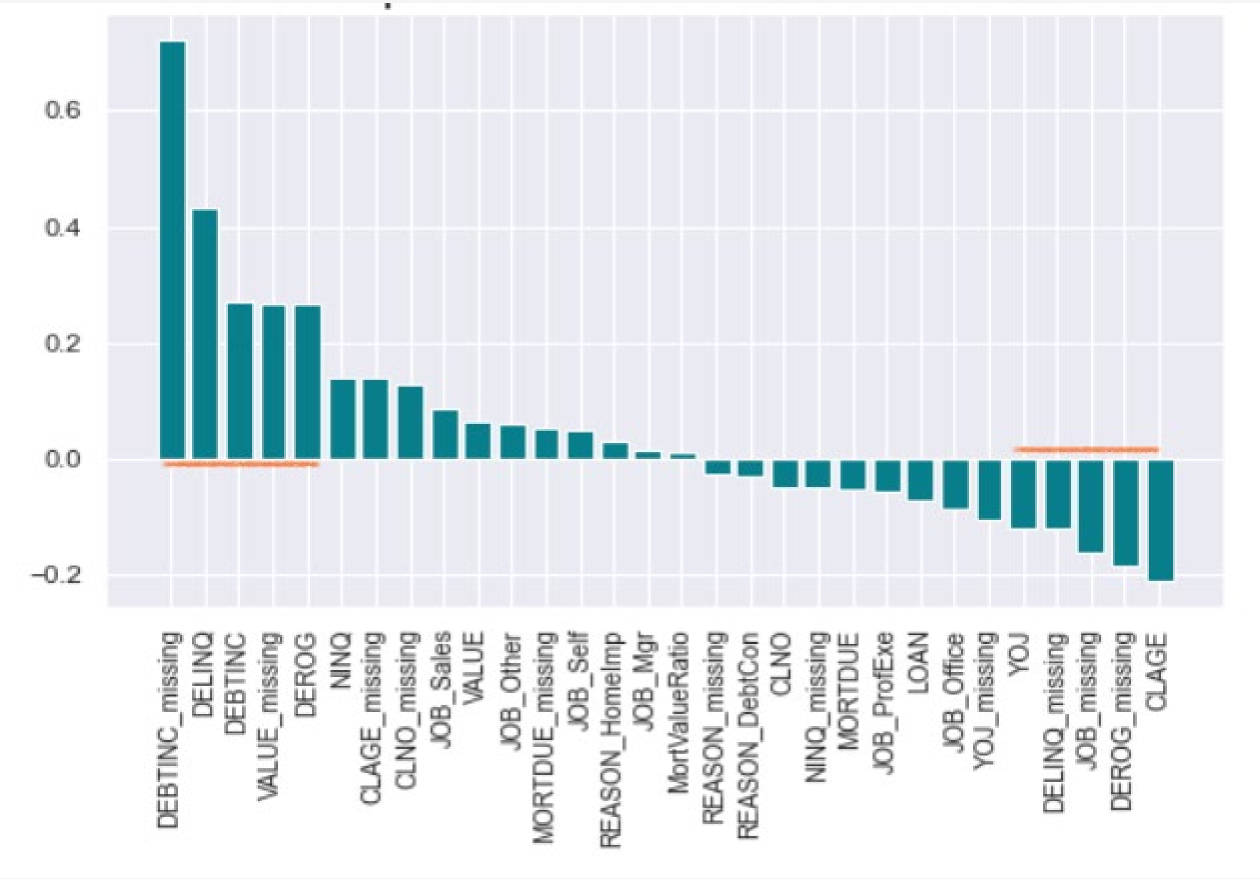
- **Goals**
  - Based on sample data, 20% of loans go bad. Prime goal was to reduce that number through data analysis and machine learning models. This was to be done while not losing too much of Good loans business
  - Needed to figure out important factors that contribute to higher probability of a given loan going Bad
  - Make recommendations based on selected model

Data Analysis and Feature Importance

Observations from Provided Data

- There were missing values in some features and some features had outlier values
- Only **Mortgage Due and Value** showed strong correlation
- DEROG, DELINQ, DEBTINC, CLAGE seemed to be contributing factors to outcome of BAD loans. This made sense because customers with bad credit history, short credit history and high Debt to income ratio can be expected to have higher chance of loan default

Observations by selected Model



In addition to provided features, selected model assigned importance to missing values of certain features also

Model Process and Performance

How was Model Performance measured?

- Out of provided 5,960 loans, we used 80% for training and 20% for validation

- Performance was measured by looking at how many Bad loans we were able to predict while keeping good number of Good loans. This was done on training and validation data.

- Most models were able to predict about 80% of Bad Loans while keeping at least 90-95% of Good loans

Selected Model Performance on Test Data

- Total Loans – 1192
- Good Loans – 964
  - Good Loans Predicted As Good Loans – 927
  - Good Loans Predicted As Bad Loans – 37 (Missed income compared to Manual)
- Bad Loans – 265
  - Bad Loans Predicted As Bad Loans – 217 (Losses Prevented compared to Manual)
  - Bad Loans Predicted As Good Loans – 48

Challenges and Further Analysis

- **Further Analysis**
  - Time when bad credit actions like DEROG or DELINQ happened was not available in provided data. In real life, bad credit 5 years back is very different from bad credit 5 months back
    - Business should try to get theses values
    - Model would need some retraining when we get those values
  - Care was taken to avoid overfitting, but more unseen data is needed for confirmation
  - It may not always be feasible to get values for all important features. Sometimes these can be for valid reasons like customer not having any debts right now, resulting in missing value for DEBTINC ratio. Model may treat this as high risk
  - We do not know if data sample was from a time period with specific economic situation (like Recession, COVID etc.) or was from different time periods

Next Steps

- **Next Steps**
  - **Business Steps**
    - Try to reduce number of missing values for important features like DEBTINC to further improve the solution
      - It has been found when DEBTINC is missing value, there is high chance of loan default
    - Customers with bad credit history like Derogatory reports or serious Delinquencies, high Debt to income ratio, short credit history have higher chance of defaulting on loan. Focus on these values for a given customer
  - **Technical Steps**
    - Deploy model so that it can be used by business
    - Monitor performance of Model once business starts using it
    - Monitor data coming through to catch any significant drift in input data
    - Project Link: https://github.com/manjit28/5511-deep-learning-project.git