# Case Overview

Concisely describe the problem and your objectives as the data analyst.

**Problem Statement**: Chess Bank deals with home loan applications and provides loan based on bank profits. The bank wants to propose a list of applicants that the bank should give loan to, by maximizing total profit from these loans. The application file for the year 2017 contains information of 250 loan applications, including loan amount and personal characteristics such as income, credit score, age, and education level.

**Objective**: As data analysts, we would like to predict whether the bank would approve the loan for the 2018 applicants and propose the list of potential applicants, by analyzing the individual characteristics and loan payment behavior of the applicants from the previous year, 2017.

# Methodology

Describe the data approach and methodology you use. Justify your choice of methodology. You should be precise, but also avoid jargon. This is a document that will be going to the CFO, so make the language appropriate for them.

As part of the analysis activity on the loan applications, we executed the following steps to draw insights from the data:

* **Data Cleaning:** As part of the initial analysis, we found that there are 19 features and 250 applicants for the year 2017.The data set is complete and balanced with no null values in the columns, therefore there was no need to perform a data cleaning.
* **Exploratory Analysis & Feature Engineering:** The mainobjective of the case is to find the list of applicants the bank should agree to give a loan based on the estimated profit on each account. We calculated the profit as the difference between the amount paid and the loan amount. We defined following new features:

1. **profit** that would take the value of ‘1’ if the profit is greater than or equal to zero and ‘0’ if the profit is less than zero.
2. **meanincome** that would be the average wage income of the loan applicants 1 and 2 years ago.
3. **li\_ratio** that would be the loan to mean income ratio of the loan applicants.

There were integer or categorical variables like ***statecode, married, educ, taxdependent*** which needed to be transformed into factor variables. The main advantage of converting this is that we can use the variables in statistical modeling where they can be implemented correctly.

On further analysis of the data set we removed some of the features which did not contribute much to the prediction result. They are namely – ***id, name, SSN, date, amt\_paid, loan\_amt, W2inc\_m1,W2inc\_m2***.The final data set now contain 14 features namely ***amt\_due,statecode,age,married,educ,taxdependent,creditscore,asset,debt,unemprate,avg\_homeprice,profit,meanincome, li\_ratio.***

* **Model Selection and Evaluation:** As part of the prediction process we used the following algorithms to derive the output:

1. **Linear Regression:** We used multiple linear regression model first to see how theindependent variables are used to predict the value of the dependent variable – profit. We divided the loans\_2017 data set into train and test with 90% values in the training set and 10% in the test set as the data set is relatively small. The training set was used to fit the model, and test set was used to evaluate the best model to get an estimation of generalization error and accuracy.

In the linear model, we got the **Accuracy** to be - **0.60**

1. **Logistic Regression:** We used logistic regression as our next predictive analysis model as it explains the relationship between the dependent binary variable well. In this case we used the cross-validation method to resample the data set.

The reported **Accuracy – 0.64**

1. **Decision Tree:** Lastly, we used the decision tree modelling technique. Decision trees works well in performing feature selection which gives a clear understanding on the features which are important deciding factors in the model. Moreover, it is easy to interpret and get insights.

In this model the reported **Accuracy – 0.72**

For all the three models we tried out different threshold values and checked the accuracy rates and the total number of applicants the bank would agree to sanction a loan and we narrowed down to the value of 0.8. For the final prediction model, we chose the decision tree method as it gave a good realistic count of applicants the bank should give a loan. For applications similar to this model, misclassifying the minority class (false negative) is a lot more expensive than misclassifying the majority class (false positive). In the context of approving a loan, the bank would have a greater probability to lose money by lending to a risky applicant who is more likely to not fully pay the loan back. This would prove to be more-costly than missing the opportunity of lending money to an applicant. Our final model predicted that the bank would approve loans for 21% of the applicants in 2018.

# Conclusion

Describe your results, including to how many people you intend to give loans, the total loan amounts, and the anticipated resulting profits. Describe your data file which reports to which individual loans should be given.

The final data file contains the loan decisions with three columns namely - ID, Name and Approve. From the loans\_2018 dataset, we could predict the number of loan applicants the bank decides to give loan. The bank would give a loan if the approve column in the file contains a ‘1’ and would not otherwise. So, out of the 250 loan applicants who applied for the loan our model predicted that the bank would give loan to 21% of the applicants in 2018 which totals to about 53 applicants. The names and the ids of the prospective applicants are included in the loan\_approval\_2018.csv file.