Introduction

Aim of the project: To predict the mortgage backed securities prepayment risk using machine leaning models.

Created by – Yasin Shah

❖ Mortgage-backed securities (MBS)

- A mortgage-backed security (MBS) is an investment similar to a bond that is made up of a bundle of home loans bought from the banks that issued them.
- In this system, the loans issued by the bank is in turn sold to investors at a discounted rate to free up the bank funds.
- These loans are sold in the form of bonds by investment banks wherein loans are grouped together according to their type and quality.
- For the investor, an MBS is as safe as the mortgage loans that back it up.

❖ Prepayment risk

- Prepayment risk is the risk involved with the premature return of principal on a fixed-income security. When debtors return part of the principal early, they do not have to make interest payments on that part of the principal.
- This means that if the loan issuer prepays the loan, the investors will stop receiving interest on those bonds.
- Hence, it is important to evaluate the prepayment risk on the MBS and thus this is the aim of our project.

❖ Data

- The data is obtained from Freddie Mac official portal for home loans.
- The size of the home loans data is (291452 x 28).
- It contains 291452 data points and 28 columns or parameters which denote different features of the data.
- Some of the noteworthy features of the dataset are: -
  - Credit score of the client
  - The maturity date of the mortgage
  - The amount or percentage of insurance on the mortgage
  - Debt to income ration of the borrower
  - Mortgage interest rate

- Prepayment Penalty Mortgage - denotes if there is any penalty levied on prepayment of loan
- Loan sequence number – denotes the unique loan ID
- The purpose of the loan
- The number of borrowers issued on the loan.
- The property type, the state in which property is and its postal code and address
- The information about the seller and service company.
- HARP indicator – denotes if the loan is HARP or non-HARP
- Interest only indicator – Denotes if the loan requires only the interest payments over the period of maturity or not.

Blueprint of project:

1. Exploratory Data Analysis (EDA)

- After obtaining the data, the first step to be performed is the EDA process. In this process, the data various operations are performed on the data to understand the data better and to get more insights from the data.
- Python libraries like pandas, matplotlib, seaborn, profile report etc. can be used to understand the data holistically.
- This helps us to know our dataset features better, the correlation of parameters, the statistical distribution of data etc.
- All this insights in turn help us to select the most suitable model for the dataset and to hyper tune it better.

2. Data cleaning and pre-processing

- In this step, the data is primarily cleaned first. This means that the rows with null values are removed, the unwanted columns are dropped from the dataset etc. so that we are able to process the data better.
- In the pre-processing step, the data is converted into consumable format for the machine learning models. The X and Y of the models are defined, labelling and encoding is performed if required and the dataset is split into training, cross-validation and test set to gauge the varied performance of the machine learning models and to evaluate their performance accordingly.

3. Model building and evaluation

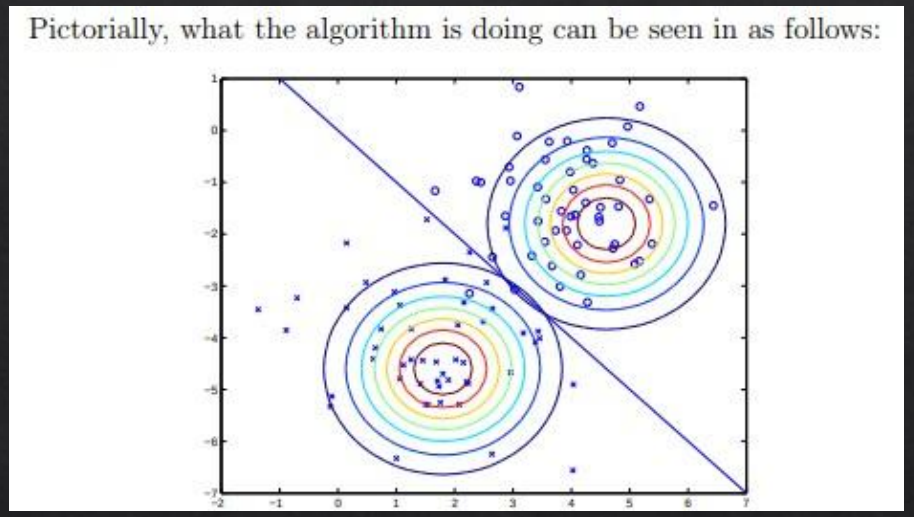
- Since this is a classification task, we will use the following algorithms

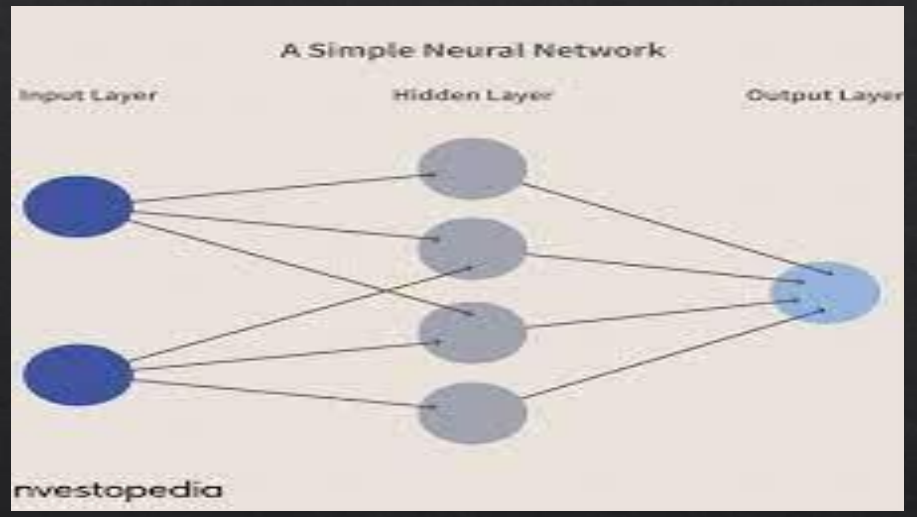
1. Logistic regression - This is a simple type of machine learning model used for classification task. At the core of this method lies the logistic function or the sigmoid function which has the formula as follows

$$P(y = 1; x, \theta) = h_x(x) = \frac{1}{1 + \exp^{-\theta^T x}}$$

1. Support Vector Machines (SVM) – SVM are types of supervised machine learning algorithms that are able to perform classification as well as regression problems. It uses a technique called the kernel trick to transform the data and then based on these transformations it finds an optimal boundary between the possible outputs. For the purpose of our project, the hyper-planes will separate the “No Prepayment” and “Prepayment” classes and we will do so by utilizing the kernel trick.

3. Gaussian Discriminant Analysis(GDA) : Gaussian Discriminant Analysis is a Generative Learning Algorithm, and in order to capture the distribution of each class, it tries to fit a Gaussian/Normal Distribution to every class of the data (given training data)separately. When using GDA to predict prepayments , we will create two Gaussian distributions, one representing the distribution of data for " No Prepayment" and one for "Prepayment".





4. Feed-Forward Neural Networks : Neural networks works by stacking and layering many neurons with activation functions together to create aa highly non-linear function. For the purposes of predicting if a prepayment occurs in the given months, we will use our dataset to train a neural network which will output the probability that a prepayment occurs.

4. Deployment

- The last stage of the project is deployment of the model on the internet using the Application Programming Interface (APIs).
- In this project we can use frameworks like Django or Flask to deploy our models after converting them to modular format.
- The models can be deployed then using platforms like Heroku, AWS or Google Cloud to use the algorithm in real-time.