Project Proposal

Erik Stenberg 3035428720
Huang Zanming 3035182081
Wong Kwan Yi 3035503192
Wong Yick Lun 3035281407

Objectives

Background of project:

Housing has been one of the biggest problems which the Hong Kong society faces. While no effective solutions seem to be available to solve it, our group aims to find out a potential strategy that HKSAR government could learn from, by looking at the government-sponsored U.S. corporate Freddie Mac, which provides open data for the public.

The mission of this company is to increase the supply of money for mortgage and hence ease up homeownership and rental housing for the U.S. citizens, which would also benefit our citizens when implemented properly in Hong Kong.

Problem of the study:

To analyse the credit performance of mortgage transactions of Freddie Mac, and potentially evaluate the feasibility of implementing Freddie Mac in Hong Kong with the help of the HKSAR government.

Project Objectives:

1) To consolidate concepts and tools that we learnt in previous/current courses by applying them in a real life dataset and situation, such as data visualisation, data mining and big data analytics etc using R, (R Core Team 2012) and the package **sparklyr** (Luraschi et al., n.d.).
2) To comprehensively execute a "Data science workflow", i.e. data preprocessing, exploratory data analysis, data modelling, model interpretation, evaluation, and data visualisation.
3) To potentially provide a strategy for the HKSAR government to alleviate the housing problem in HK.

Description of data

Source of Data

The dataset utilised in this project is provided by The Federal Home Loan Mortgage Corporation (FHLMC), also known as Freddie Mac. Freddie Mac is a government-sponsored private corporation which aims at providing liquidity, stability and affordability to the U.S. housing market by purchasing residential mortgage loans from the lenders on the secondary mortgage market and consequently sell them to other investors on the market, according to their website, (see Freddie Mac 2018a). By doing this, Freddie Mac maintains a stable money flow to the mortgage loan lenders on the market in pursuits of easier homeownership and rental housing for the U.S. citizens.

The liability of the dataset can be guaranteed as Freddie Mac is regulated by the Federal Housing Finance Agency (FHFA), an independent federal agency with expanded legal and regulatory authority and power of several departments of the U.S. Federal Government, such as the Federal Housing Finance Board (FHFB), the Office of Federal Housing Enterprise Oversight (OFHEO) etc. Regardless of the trustworthiness of the dataset, incompleteness and errors are expected.

Method of data collection

In hopes of increasing transparency, Freddie Mac has disclosed monthly loan-level credit performance data on a portion of fully amortising fixed-rate mortgages that the company purchased or guaranteed from 1999 to 2017, which could be viewed and downloaded freely on their official website.

Background information of the data

The Single Family Loan-Level Dataset (the dataset) originally covers approximately 26 million fixed-rate mortgages. Albeit the huge data size, our group will focus on the data between 2007 to 2016 due to limited hardware power of the machines such as memory and storage.

The dataset has two parts: loan origination data and monthly performance data. In the case where the loan is in the origination part but not the performance part, the loan was paid off before the first cycle begins or in the month of origination. The variables in the dataset will be discussed in later sessions.

Description of variables:

Origination Dataset

In the origination data file, there are a total of 26 variables. The table below describes the variables in the dataset.

	Variable Name	Description	Type
1	Credit Score	Represents the borrower's creditworthiness and indicates the likelihood he/she will timely repay future obligation	Numeric
2	First payment date	Date of the first scheduled mortgage payment.	Numeric
3	First time homebuyer flag	Indicates whether the Borrower is an individual who (1) is purchasing the mortgaged property, (2) will reside in the mortgaged property as a primary residence and (3) had no ownership interest (sole or joint) in a residential property during the three-year period preceding the date of the purchase of the mortgaged property.	Categorical Y=Yes N=No 9=Not Available
4	Maturity date	The date in which the final monthly payment on the mortgage is scheduled to be made as stated on the original mortgage note.	Numeric
5	Metropolitan Division	Indicates the metropolitan in which the mortgaged property is located.	Numeric

	Variable Name	Description	Type
6	Mortgage Insurance Percentage (MI%)	The percentage of loss coverage on the loan, at the time of Freddie Mac's purchase of the mortgage loan	Numeric 1% - 55% 000 = No MI 999 = Not Available
7	Number of units	whether the mortgage is a one-, two-, three-, or four-unit property.	Numeric
8	Occupancy Status	whether the mortgage type is owner occupied, second home, or investment property.	Categorical P = Primary Residence I = Investment Property S = Second Home 9 = Not Available
9	Original Combined Loan-To-value (CLT)	0% - 200% 999 = Not Available	Numeric
10	Original Debt-to-Income (DTI) Ratio	Ratios greater than 65% are indicated that data is Not Available.	Numeric
11	Original UPB	The UPB of the mortgage	Numeric
12	Original loan to value (LTV)	Ratios below 6% or greater than 105% will be disclosed as "Not Available," indicated by 999.	Numeric
13	Original Interest Rate	The original note rate as indicated on the mortgage note	Numeric
14	Channel	R=Retail B=Broker C=Correspondent T=TPO Not Specified 9=Not Available	Categorical
15	Prepayment penalty mortage (PPM) Flag	A mortgage with respect to which the borrower is, or at any time has been, obligated to pay a penalty in the event of certain repayments of principal.	Categorical Y=PPM N=Not PPM
16	Product type	Denotes that the product is a fixed-rate mortgage.	Categorical
17	Property State	Indicating the state or territory within which the property securing the mortgage is located.	Categorical
18	Property Type	Denotes whether the property type secured by the mortgage is a condominium, leasehold, planned unit development (PUD), cooperative share, manufactured home, or Single Family home.	Categorical
19	Postal Code	The postal code for the location of the mortgaged property	Numeric
20	Loan sequence number	Unique identifier assigned to each loan.	Categorical

	Variable Name	Description	Type
21	Loan Purpose	P = Purchase C = Cash-out Refinance N = No Cash-out Refinance 9 =Not Available	Categorical
22	Original Loan Term	A calculation of the number of scheduled monthly payments of the mortgage based on the First Payment Date and Maturity Date	Numeric
23	Number of borrowers	Number of borrowers of the loan.	Numeric
24	Seller Name	The name of the seller.	Categorical
25	Servicer Name	Name of servicer.	Categorical
26	Super conforming flag	For mortgages that exceed conforming loan limits with origination dates on or after 10/1/2008 and settlements on or after 1/1/2009	Categorical

Source: Freddie Mac (2018b)

Monthly Performance Dataset

In the monthly performance dataset, there are a total of 24 variables. The table below describes the variables in the dataset.

	Variable Name	Description	Type
1	Loan Sequence Number	Unique identifier assigned to each loan.	Categorical
2	Monthly reporting period	The as-of month for loan information contained in the loan record.	Numeric
3	Current Actual UPB	Reflects the mortgage ending balance as reported by the servicer for the corresponding monthly reporting period.	Numeric
4	Current Loan Delinquency Status	A value corresponding to the number of days the borrower is delinquent	Categorical
5	Loan Age	The number of months since the note origination month of the mortgage.	Numeric
6	Remaining months to legal maturity	Unique identifier assigned to each loan.	categorical
7	Repurchase flag	Indicates loans that have been repurchased or made whole (not inclusive of pool-level repurchase settlements).	categorical
8	Modification Flag	For mortgages with loan modifications, indicates that the loan has been modified.	Categorical
9	Zero Balance Code	A code indicating the reason the loan's balance was reduced to zero.	Categorical
10	Zero Balance Effective Date	The date on which the event triggering the Zero Balance Code took place.	Date

	Variable Name	Description	Type
11	Current Interest Rate	Reflects the current interest rate on the mortgage note, taking into account any loan modifications.	Numeric
12	Current Deferred UPB	The current non-interest bearing UPB of the modified mortgage.	Numeric
13	Due Date of Last Paid Installment (DDLPI)	The due date that the loan's scheduled principal and interest is paid through, regardless of when the installment payment was actually made.	Date
14	MI Recoveries	Mortgage Insurance Recoveries are proceeds received by Freddie Mac in the event of credit losses.	Numeric
15	Net Sales Proceeds	The amount remitted to Freddie Mac resulting from a property disposition.	Categorical,Numeric
16	Non-MI Recoveries	Non-MI Recoveries are proceeds received by Freddie Mac based on repurchase/make whole proceeds, non-sale income such as refunds (tax or insurance), hazard insurance proceeds, rental receipts, positive escrow and/or other miscellaneous credits.	Numeric
17	Expenses	Expenses will include allowable expenses that Freddie Mac bears in the process of acquiring, maintaining and/ or disposing a property (excluding selling expenses, which are subtracted from gross sales proceeds to derive net sales proceeds).	Numeric
18	Legal Costs	The amount of legal costs associated with the sale of a property (but not included in Net Sale Proceeds).	Numeric
19	Maintenance and preservation costs	The amount of maintenance, preservation, and repair costs, including but not limited to property inspection, homeowner's association, utilities, and REO management, that is associated with the sale of a property (but not included in Net Sale Proceeds).	Numeric
20	Taxes and Insurance	The amount of taxes and insurance owed that are associated with the sale of a property (but not included in Net Sale Proceeds)	Numeric
21	Miscellanesous Expenses	Miscellaneous expenses associated with the sale of a property (but not included in net sales proceeds)	Numeric
22	Actual Loss Calculation	Actual Loss of the loan	Numeric

	Variable Name	Description	Type
23	Modification Cost	The cumulative modification cost amount calculated when Freddie Mac determines such mortgage loan has experienced a rate modification event.	Numeric

Source: Freddie Mac (2018b)

Quality of the data and data preparation:

1) Many empty or N/A values in the monthly performance data set. The Dataset is sparse in nature. 15 out of 23 total columns are mostly empty or not avalible. Some columns such as Repurchase Flag, Modification Flag, Zero Balance Code, and Zero Balance Effective Data are left empty on purpose, as these data are only avalible in the loan termination months. Also columns like Legal Costs, Tax and Insurance, Miscellaneous Expenses are left empty as the data are only applicable when loan ends with foreclosure or REO Disposition. However other columns such as Actual Loss Calculation are left empty without expanation. Due to the lack of data, there will be certain limitations in future analysis.
2) Many of the variables in the data are coded in ways that may not be optimal for further analysis. We will therefore recode many of the variable into forms that are easier to handle. Below are a few examples that represents most of the variable recodings done.

Variable	Current range	New Range	Example Manipulation
First Time Homebuyer Flag	Y=YES N=NO Space(1) for Unknown/miss	Y = 1 N = 0	`data %>% mutate(flag= if_else(fthb=="Y", 1, if_else(fthb=="N",0,NA)))`
Credit Score	301-850 Space(3) = Unknown, if CS < 301 or CS > 851	301-850	`data %>% filter(CS %in% c(301:850))`

3) Many binary variables are coded as 'YES' and 'NO' and spaces for unknown, while in R those variables should be recoded into numerical `1` for 'YES', `0` for 'NO' and `NA` for the missing values. The credit score variable serves as another good example of when we need to be careful. There is no way of knowing whether a missing value in credit score means high (below 301) or high (above 850). Therefore, it is essential that we remove all rows with missing value on credit score when doing any analysis relate to it. Any other way of dealing with these missing values will certainly introduce bias. Exstensive data wrangling, manipulation and creation of new variables is inevitable since people have different intentions of usage and use different softwares.

Study Plan

Since, the data set have hugh number of observations and variables. We will handle it in sparklyr.

Data Visualisation

We plan to do the following Data visualizations on the dataset:

1) Create an animated geographical map using **leaflet** (Cheng, Karambelkar, and Xie 2017) to show the location of the houses under mortgage either using the property state variable or the postal code variable in the origination data set. We can also investigate the differences between the different states through the map.

2) Create a treemap using the loan purpose variable in the origination data set to analyze the loan purpose.

3) Create time series plots on the current interest rate variable in the monthly performance data set and potentially other suitable variables. We can use these plots to investigate if there are any trends or patterns in the time series.

Modelling

We plan to adapt the following modeling method in the project:

1) To see if it is possible to fit a linear regression with credit score being the response variable and the geographical location and other non-mortage-realted variables as the predictor variables. If so, we can use the regression model to predict the credit score. However, there are certain problems that might occur. Firstly, we need to ensure if there are any linear relationship between credit scores and the other variables. Second, we need to check for multicollinearity. If there is, we cannot fit a linear regression model to it and thus, we will have to look for other regressions model that might be a better fit to this dataset. We will start simple and work our way up to more sophisticated algorithms, and fine tune them accordingly, splitting the data set into training and testing parts.

2) Using the time series we will create, we can follow up by doing a time series forecasting on the current interest rate variable in the monthly performance dataset. Firstly, we can see if it is possible to fit any forecasting model such as ARIMA, ETS, etc. Secondly, we can test the accuracy of the forecasted values against the actual values in the dataset.

Referenecs

Cheng, Joe, Bhaskar Karambelkar, and Yihui Xie. 2017. *Leaflet: Create Interactive Web Maps with the Javascript 'Leaflet' Library.* https://CRAN.R-project.org/package=leaflet.

Freddie Mac. 2018a. "Freddie Mac." http://www.freddiemac.com.

———. 2018b. *Single Family Loan-Level Dataset General User Guide.* Freddie Mac.

Luraschi, Javier, Kevin Kuo, Kevin Ushey, JJ Allaire, and The Apache Software Foundation. n.d. *Sparklyr: R Interface to Apache Spark.* http://spark.rstudio.com.

R Core Team. 2012. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.