David Sacco

Fairfield University

Capstone Project

Business Analytics Capstone Project Proposal

2019 June 11

Subject Matter Expert: Robert Sacco

MSBA Capstone Advisor: Dr. Yasin Ozcelik

**Project Background/Motivation**

The motivation for this is to start to bring people away from using Excel based graphs and charts, as Excel is very clunky and outdated compared to other programs that are not being utilized by big companies because they are afraid to make the transition. R Shiny is at the forefront of data visualization, and I was not able to have the opportunity to use it in the Business Analytics Program, but major companies like to see that potential employees have a background in using R, even if it is for visualizations only.

Excel is very slow sometimes when handling hundreds of thousands of rows of data that has a vast number of attributes. The power that R has as a statistical programming language is underutilized by major companies, and I would like to show a team in the mortgage banking sector at JP Morgan Chase the advantages of using R over Excel.

**Problem Statement/Statement of Work**

The project that I intend to create and deliver includes a data warehouse containing years of public use data from Fannie Mae and Freddie Mac that will serve as a backbone for a visualization application using R Shiny. The purpose of the application is to create a simple way for the end user to use a single URL in their web browser and filter the data by specific criteria and have it show them a high overview of charts that are useful to them.

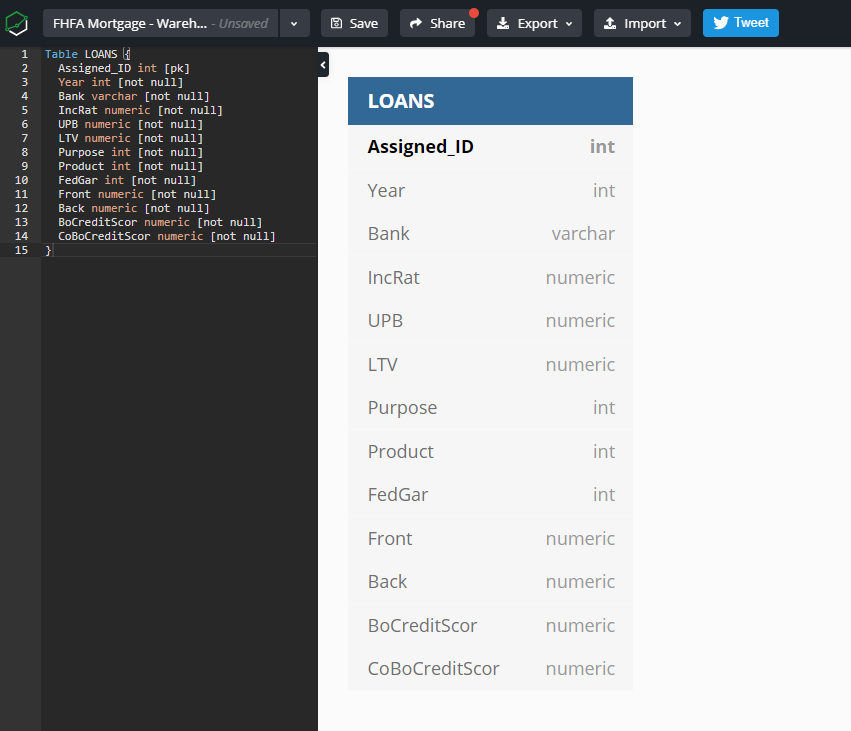
The database is going to be hosted on AWS cloud using PostgreSQL as the infrastructure of the data warehouse. I have never used PostgreSQL before, but I am aware that there are forums and discussion boards for help when I am stuck. From my understanding, there are minor differences from using MySQL and Oracle, which I have experience in using SQL. I will be able to troubleshoot any SQL errors that I have through those forums and discussion boards.

I have to also create some sort of analytical model that will be used with the data. I will most likely use a logistic regression model, or any other models that are suggested to create some sort of model for identifying if a loan is going to be approved or not. I am open to suggestions for this, as I already have base code for this saved in GitHub, so I would have to do some data cleaning and then some tinkering with the data and I will be able to create a model.

I will be using my personal computer as my hub for running Python and R, as my computer has more capabilities with RAM and computer processing for major computations, so AWS is not necessary for me.

**Preliminary Results/Exploratory Data Analysis**

ERD for the Data Warehouse:



Speaking to the ERD, I am going to create only one table for the data warehouse because one record is one unique loan. Please see below for the data dictionary for each attribute:

**Year** - Year

**Assigned ID –** Unique Record ID (Not Actual Loan Number)

**Bank –** Name of Federal Home Loan Bank District

**IncRat –** Borrower Income Ratio

**UPB –** Acquisition Unpaid Principal Balance in Whole Dollars

**LTV –** Loan to Value Ratio at Origination

**Purpose –** Loan Purpose [1 = purchase (1 multiplier); 2 = refinancing (1.61 multiplier); 3 = second mortgage (2.23 multiplier); 4 = new construction (1 multiplier); 5 = rehabilitation (1.707 multiplier)]

**Product =** Product Type [1 = Fixed Rate (0.879 multiplier); 2 = ARM (1 multiplier); 3 = Balloon (1.24 multiplier); 4 = GPM/GEM (1.24 multiplier); 5 = Reverse Annuity Mortgage (1.24 multiplier); 6 = other (1.24 multiplier)]

**FedGar =** Federal Guarantee [0 = No Federal Guarantee; 1 = FHA; 2 = VA; 3 = FMHA-Guaranteed Rural Housing Loan; 4 = HECMs; 5 = Title1-FHA]

**Occup** = [1 = Principal residence/Owner-occupied (1 multiplier); 2 = second home (1.304 multiplier); 3 = investment property (rental) (2.141 multiplier)]

**Front =** Front-end Ratio

**Back =** Back-end Ratio

**BoCreditScor** = Borrower Credit Score [1 = <620 (7.426 multiplier); 2 = 620 to < 660 (4.81 multiplier); 3 = 660 to < 700 (2.85 multiplier); 4 = 700 to < 760 (1.76 multiplier); 5 = 760 or greater (0.743 multiplier); 9 = missing]

**PropType** = [PT01 = Single Family Detached (1 multiplier); PT02 = Deminimus PUD (1 multiplier); PT03 = Single Family Attached (1 multiplier); PT04 = Two family (1 multiplier); PT05 = Townhouse (1 multiplier); PT06 = Low-Rise Condominium (1.146 multiplier); PT07 = PUD (1 multiplier); PT08 = Duplex (1 multiplier); PT09 = Three family (1.146 multiplier); PT10 = Four family (1.146 multiplier); PT11 = High-rise condominium (1.146 multiplier); PT12 = Manufactured Home (1.93 multiplier)]

**Coop** = [1 = Yes (0.345 multiplier); 2 = No (1 multiplier)]

**Planning Timeline with Milestones**

**Project Proposal Approval**: June 11

**Data Warehouse ERD**: June 7

**R Shiny Layout**: June 14

**Database Completion**: June 13

**R Shiny App Completion**: June 17

**Analytics Completion**: June 20

**Summary of Expected Results**

The expected results from the project is to have a fully functioning data warehouse that will be hosted on the cloud connected to an R Shiny application that will run on a server in the cloud, and ultimately create a couple of models that will help predict if certain loans will be approved or not.

I will create a script to test the data quality for the data warehouse to ensure a level of data quality when querying the warehouse. I will meet with the team at JP Morgan Chase to ensure that the data that they are looking at meets their expectations, and I will create the analytical section to fulfill the capstone requirement.

**Follow-Up Questions from Panel**

1. The client does not have any requirements for the database system, as they will not be able to get into the database. I intend to keep that privilege to myself only so I know no one else is messing with the database. It will be used for housing the data that I have, and serve as a database server that the R Shiny app will query when generating visuals. It is initially for internal use only, but since the data is public, it can be used for external purposes if it has to be repurposed. I only intend to share this connection with the team that I am creating it for.
2. Please see the **Preliminary Results /Exploratory Data Analysis** section of the proposal for all of the attributes that I will include in the database. My main point of contact with the JP Morgan Chase Mortgage Banking team is Rob Sacco. I have sat down with him and discussed the major attributes that he and his team look at when generating the visuals that he has provided to me. Those visuals I cannot share with the panel, as there is data on there that is confidential.
3. For the ERD, please see the **Preliminary Results/Exploratory Data Analysis** section of the proposal.
4. The specific purpose that the client is going to use the app for is data visualization. They will then analyze the data after they have created the visuals. Their analysis is a bit different than data analytics, as they are more focused on certain aspects of year-over-year data for the loans. I am unsure if the R Shiny app can be a part of a story line, as I want to create something that is easy and straight forward to use for people that have been using Excel for the past 20 years so I do not overwhelm them. I can use Tableau instead, but R Shiny is more powerful in my opinion. I would like to use this project in future job applications and R is a more sought-after skill if I can create a functioning application connected to a database that has visuals tailored to a certain group of people.
5. I will create a user manual for the users, as this will be a final touch that I will put together for the team. I have some experience in creating these User Manuals, so I will create something at the end of the creation of the app. I will format it for both a PDF and a Word document for the end-users.
6. For the analytics, I have base code from Dr. Tao and Dr. Huntley’s classes that I can use to import the data sets and do some exploratory analysis, data cleaning, and ultimately build a model that will determine if a loan will be approved or not depending on certain attributes. I will do this as a last step, as I have base code for this already, so I will have to do some tweaking and modifications to the code to make it work the way I would like it to. I will do this in Python to showcase that I can use SQL, R, and Python.