# Industrial Course

# Data Science and Big Data Analytics

# Use Case – Workflow 1

FACULTY OF COMPUTERS AND INFORMATION

CAIRO UNIVERSITY

# Purpose and references

- This lab allows students to apply what they have learned from the analytical methods and tools to a big data problem using the Analytics Lab Environment.
- References used throughout the labs are located in your ***Student Resource Guide Appendix***. See the Appendix for:
    - http://www.ffiec.gov/hmda/

# Tasks

- Tasks you will complete in this lab exercise include:

  □ Explore the big data set provided and prepare the data for analysis

  □ Assess data quality, outliers and training sets

  □ Conduct model selection, code, execute and score the model

  □ Use R and PSQL statements during your analysis of big data

  □ Create a narrative summary of your findings, using the methods covered earlier in this module

# Scenario

- A financial planning company, FPC would like to expand the set of services they offer by creating an online site for loan advice. Potential home loan borrowers can enter information about their personal finances and the kind of home loan they want, and the site will return the probability of getting such a loan, along with some general advice about how to increase their likelihood of success. For example, the advice could be: "Increase the down payment so as to decrease the loan amount by X dollars"; or "Consider a home in the price range Y"; "Are you eligible for a particular type of loan?", or "Can you add a co-signer to the loan?".
- The company hopes this online service will be a lead-in for customers to come to FPC for more focused, personal financial planning to achieve their life goals. FPC would also consider partnering with a real estate broker to showcase houses to potential homebuyers. FPC realizes that the customers are looking for fast responses and the online-service must provide an answer within 45 seconds. FPC plans to enter into a service level agreement with the partner websites such as those managed by the real estate brokers.
- Ideally, the model behind this advice site can give reasonable, grounded predictions. Of course, the site cannot ask applicants to fill out an entire loan application and the sensitive data it contains, such as credit scores, employment history, or existing debt. The FPC project stakeholders want to stick to basic, easy to enter information such as applicant income, loan type, loan size, and the location of the property (ZIP code).
- They recognize that a model with only that information can only give general advice, rather than truly precise predictions. We have a set of data that can support this approach, and allow making predictions based on the information above.

# Issues to Address

- A number of issues came up during the kick-off meeting for the project:

  1. Should there be one big model, or separate models for different types of loans?

  2. Someone in the group wondered if personal demographic information (sex, gender, and ethnicity) would improve the prediction. The others are hesitant about the idea of asking such questions on the site, but agreed to explore whether knowing that information would improve the model.

  3. Someone else offered the opinion that giving the users raw probabilities would not be meaningful to them. She suggested that the model should set thresholds, and deliver qualitative messages instead, such as the following:

  □ If the model reports that the probability of getting a loan were greater than 75%, then the system would send the user a message such as: “Congratulations! You have a very good chance of getting your loan!”

  □ For probability less than 50%: “Sorry. Looks like the chances aren’t so good,” with a link to FPC’s advice page.

  □ For probability between 50-75%: “Your chances aren’t the strongest. Come talk to us about developing a plan to improve your chances of getting financing.”

This work led the group to a metric for measuring model performance. Of the people who score > 75%, do more than 75% of them actually get a loan? Likewise for people who score less than 50%, how many of them are actually get loan? Also, how many people in the general population get each message? For instance, does the entire population score more than 75%?

# Data scientist role

- Your goal, as the data scientist on this project, is to answer the following questions:

  1. Would it be more effective to develop different models or one model? Why? If different models, focus on a single one for the initial study.

  2. Should we ask for personal demographic information, or can we build a good enough model without it?

  3. How accurate is the model, in terms of the thresholds that the stakeholders set in their discussion (75% and 50%)? What is the coverage of the threshold regimes?

  4. Provide suggestions for the kind of general advice FPC can put on their advice page.

# Considerations for developing an Analytic plan

- Consider the scope of the data you will need to include in the analysis, and filters you may need to set to construct the data set for your analysis.
- Consider the types of models best suited to perform the analysis needed for the new website engine. Does this scenario represent a classification, clustering, or prediction problem?
- Examine the distribution of data, such as loan data for home improvement, home purchase, and refinancing loans, to identify the influences on how you will select and create the model
- Look at creating several models and compare them in terms of ROC/AUC, or other performance metrics.
- Find ways to examine how robust the model is with the help of a confusion matrix or similar diagnostic technique.
- Give thought to how you would portray this information to business stakeholders as well as an analytical audience.
- Consider the Service Level Targets that FPC can offer to their end users when they score the model with their inputs
- Consider Service Level Targets that you can provide to FPC in terms of computational resources required for model generation and validation.
- Provide some suggestions for the kind of general advice FPC can put on their advice page, based on the results from your modeling exercise. For instance, mention the types of things an applicant can do to increase their likelihood of success when applying for a loan on the website.

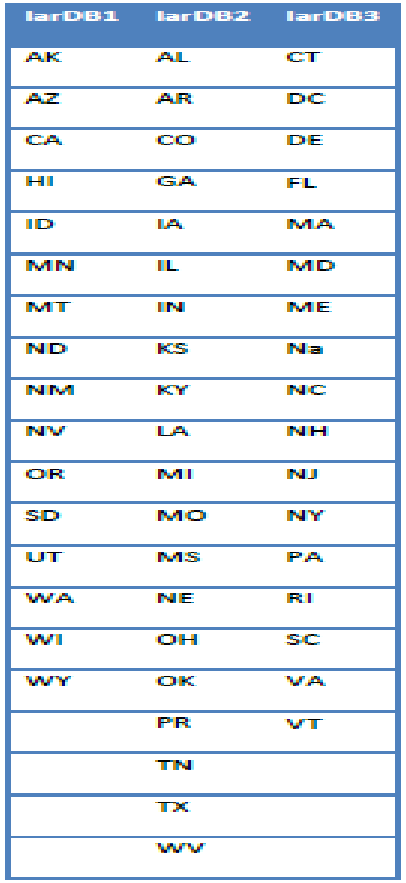
# Workflow 1

1- The data for this lab is the housing loan database assembled by federal agencies pursuant to the Home Mortgage Disclosure Act (HMDA).

This database identifies the census tract location of almost every housing loan and housing loan application made in the United States each year.

The data provided for analysis in this lab is an extract for the year 2010.

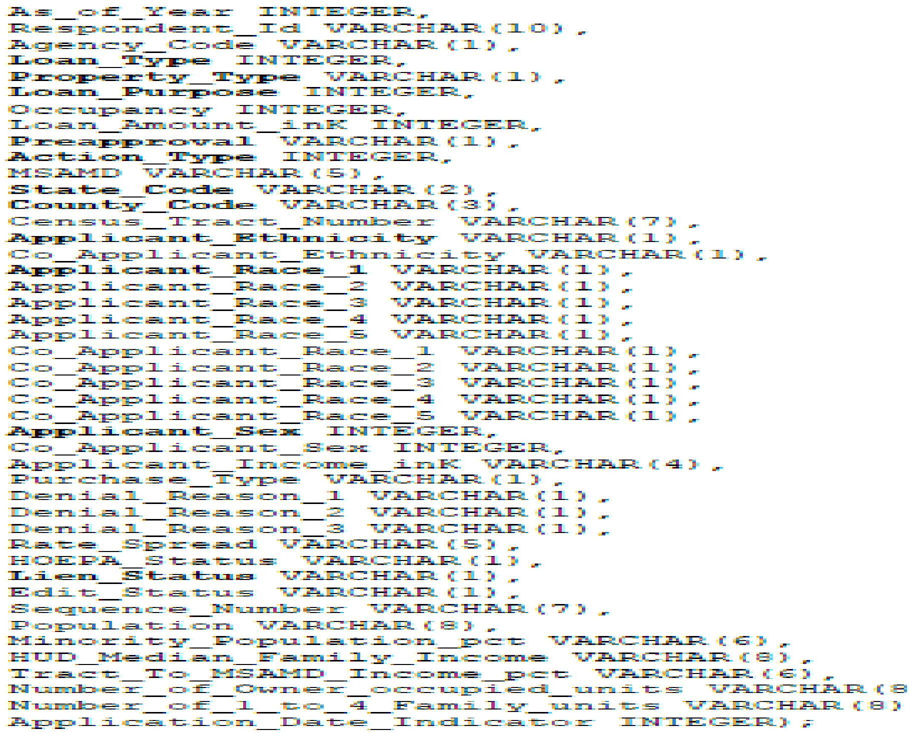
The data is organized in three database tables larDB1,larDB2,larDB3 (in database “hmdalab”) for different states as follows



| larDB1 | larDB2 | larDB3 |
|---|---|---|
| AK | AL | CT |
| AZ | AR | DC |
| CA | CO | DE |
| HI | GA | FL |
| ID | IA | MA |
| MN | IL | MD |
| MT | IN | ME |
| ND | KS | Na |
| NM | KY | NC |
| NV | LA | NH |
| OR | MI | NJ |
| SD | MO | NY |
| UT | MS | PA |
| WA | NE | RI |
| WI | OH | SC |
| WY | OK | VA |
| | PR | VT |
| | TN | |
| | TX | |
| | WV | |

# Workflow 1 - Continued

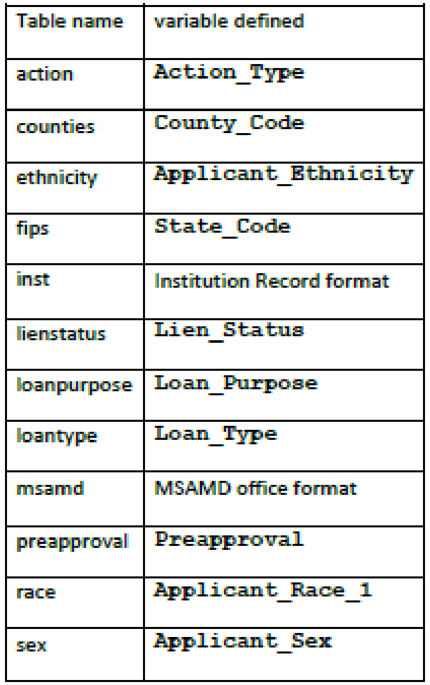
2- The tables provide the HMDA Loan Application Registration (lar) details and they have the following structure:



```
As_of_Year INTEGER,
Respondent_Id VARCHAR(10),
Agency_Code VARCHAR(1),
Loan_Type INTEGER,
Property_Type VARCHAR(1),
Loan_Purpose INTEGER,
Occupancy INTEGER,
Loan_Amount_inK INTEGER,
Preapproval VARCHAR(1),
Action_Type INTEGER,
MSAMD VARCHAR(5),
State_Code VARCHAR(2),
County_Code VARCHAR(3),
Census_Tract_Number VARCHAR(7),
Applicant_Ethnicity VARCHAR(1),
Co_Applicant_Ethnicity VARCHAR(1),
Applicant_Race_1 VARCHAR(1),
Applicant_Race_2 VARCHAR(1),
Applicant_Race_3 VARCHAR(1),
Applicant_Race_4 VARCHAR(1),
Applicant_Race_5 VARCHAR(1),
Co_Applicant_Race_1 VARCHAR(1),
Co_Applicant_Race_2 VARCHAR(1),
Co_Applicant_Race_3 VARCHAR(1),
Co_Applicant_Race_4 VARCHAR(1),
Co_Applicant_Race_5 VARCHAR(1),
Applicant_Sex INTEGER,
Co_Applicant_Sex INTEGER,
Applicant_Income_inK VARCHAR(4),
Purchase_Type VARCHAR(1),
Denial_Reason_1 VARCHAR(1),
Denial_Reason_2 VARCHAR(1),
Denial_Reason_3 VARCHAR(1),
Rate_Spread VARCHAR(5),
HOEPA_Status VARCHAR(1),
Lien_Status VARCHAR(1),
Edit_Status VARCHAR(1),
Sequence_Number VARCHAR(7),
Population VARCHAR(8),
Minority_Population_pct VARCHAR(6),
HUD_Median_Family_Income VARCHAR(8),
Tract_To_MSAMD_Income_pct VARCHAR(6),
Number_of_Owner_occupied_units VARCHAR(8
Number_of_1_to_4_Family_units VARCHAR(8)
Application_Date_Indicator INTEGER);
```

# Workflow 1 - Continued

3- All the required codes for the modeling exercise are made available in different tables as detailed below:



| Table name | variable defined |
|---|---|
| action | Action_Type |
| counties | County_Code |
| ethnicity | Applicant_Ethnicity |
| fips | State_Code |
| inst | Institution Record format |
| lienstatus | Lien_Status |
| loanpurpose | Loan_Purpose |
| loantype | Loan_Type |
| msamd | MSAMD office format |
| preapproval | Preapproval |
| race | Applicant_Race_1 |
| sex | Applicant_Sex |

# Workflow 1 - Continued

- Property type is not coded in a table, but has code definitions as follows:

1: 1 to 4 family

2: Manufactured housing

3: Multi-family

- Occupancy = 1 indicates owner occupied housing (our focus of analysis)

**4. For your analysis you are required to select**

**a. A single state**

**b. Occupancy = 1**

**c. Property_Type = 1**

**d. Action_Type <= 4**

# Workflow 1 - Continued

5- Extract data from the "lar" table (with the conditions in step 4) and create a table with the following variables:

| | | |
|---|---|---|
| **Loan_Type VARCHAR(20),** | **Applicant_Ethnicity VARCHAR(25),** | HOEPA_Status VARCHAR(1), |
| **Loan_Purpose VARCHAR(25),** | Co_Applicant_Ethnicity VARCHAR(1), | **Lien_Status VARCHAR(25),** |
| Loan_Amount_inK INTEGER, | **Applicant_Race_1 VARCHAR(25),** | Minority_Population_pct VARCHAR(6), |
| **Preapproval VARCHAR(25),** | **Applicant_Sex VARCHAR(25),** | HUD_Median_Family_Income VARCHAR(8), |
| **Action_Type VARCHAR(25),** | Applicant_Income_inK VARCHAR(4), | Tract_To_MSAMD_Income_pct VARCHAR(6), |
| **County_Name VARCHAR(50),** | Rate_Spread VARCHAR(5), | Number_of_Owner_occupied_units VARCHAR(8) |

The highlighted variables must be expanded to the values corresponding to the codes in the "lar" table.