Mortgage Pooling Combinatorial Optimization Challenge

Mission

Given the background and the CSV dataset, what useful information would you predict? How would you analyze the data?

- We want to see what visualization tools you use.
  - What BI tools are used and extent of knowledge in that tool. AWS Quicksight is a huge plus.
    - Quicksight = 10
    - All other BI tools = 5
  - We want to know what data structure you use to save the data. What columnar storage database you used and how you're loading data into the BI tool. Athena/Redshift are a bonus.
    - BI Tools:
      - Athena/Redshift = 10
      - All other
    - Data formats:
      - Parquet - 10
      - Json/CSV - 5
- We want to see what questions you ask and what they help us answer.
  - What graphs and widgets are created?
  - You will be rated on how many questions (min 5 questions) you can ask and how it helps us solve the mortgage pooling problem, i.e. which pool is the best to put these loans into.
- We want to know how you represent the data.
  - How was the data visualized? i.e. using graphs, tables, pivot tables, etc. e.g. If you used a scatter plot when a linear graph was clearly the better option, that might penalize your score.

Background

A given bank might lend out a couple hundred mortgages to various home buyers. The bank typically wants to liquidate these loans by packaging and selling them to government entities such as Fannie Mae or Freddie Mac at a small markup. In this way, the bank not only makes money on fees and the markup, but also frees up these assets for providing more loans, and repeating the cycle.

Fannie Mae and Freddie Mac have pre-defined pools for creating their mortgage-backed securities, each of which has a price they are willing to pay to the bank for that loan. E.g. FNMA 3.0 December shows a price of 101.38, while FHLCMC 3.0 December shows a price of 101.44, meaning that Fannie Mae's markup for their December 3% coupon is 1.38% and Freddie Mac's is 1.44%. It is in the bank's best interest to sell each loan for the highest price possible (i.e. sell into the pool which offers the highest price). Among these two pools, the bank should allocate the loan to FHLCMC 3.0 December.

The price that a given loan can receive is further affected by the servicing rights owner. Additional restrictions and constraints exist in each pool and servicing combination, such that maximizing the total price obtained across all pool allocations is a complex optimization. The following sections will outline the optimization goal and constraints.

Data Description

There are three tabs of data.

The *Eligible Pricing Combinations* tab shows the available pool options j and servicing options k for each loan i, and the corresponding price. Each row is a unique eligible pool & servicing option for a loan. For example, in rows 34-36, the pool option is #126 for all three rows, but the servicer is different, hence different prices P_{ijk}.

The *Loan Data* tab contains relevant info for each loan i, including:

- LoanID – unique identifier for loan i
- LoanAmount – value of the loan, L_i
- FICO – credit score
- DTI – debt-to-income ratio
- HighBalFlag – identifies whether a loan is high balance or not
- PropOcc – Property Occupancy type indicator. Three types: Primary, Secondary, NOO
- PropState – Property State indicator. Assume all 50 states possible.
- PropType – Property Type indicator. Thirteen types: 2 Unit, 3 Unit, 4 Unit, Co-op, CondoHi, CondoLo, CondoMid, DetCondo, Manu, Modular, NonWarrantCondo, SFR_At, SFR_Dt
- Purpose – Loan purpose indicator. Three types: Purchase, Cashout, Rate/Term

The *Pool Option Data* tab contains type information for each pool j, including:

- Pool Type – indicates whether the pool is Single-Issuer or Multi-Issuer
- Pool Balance Type – indicates whether pool is Standard Balance or High Balance

- Agency – indicates whether pool is from Freddie Mac or Fannie Mae