**# Background:**

**Mortgage delinquency** refers to a **borrower's failure to make timely mortgage payments**. It is a significant issue in the mortgage industry because it can have a **negative impact** on both the borrower and the lender. For the *borrower***,** delinquency can *lead to financial stress*and, in severe cases, *result in foreclosure*. For the *lender*, it can mean a *loss of income* and an *increase in the number of non-performing loans on their books*.

Studying mortgage delinquency and its relationship with other variables, such as income, employment, and credit score, is important because it can help lenders identify risk factors that contribute to delinquency.

This information can be used **to develop strategies** to **minimize** the **risk of delinquency and improve the overall performance of their mortgage portfolio**. Additionally, understanding the factors that contribute to mortgage delinquency **can help policymakers develop programs** and policies **to support borrowers and improve financial stability** in the housing market.

**# Purpose:**

Our group decided to conduct an experimental analysis to find out if there were other factors that impact mortgage delinquency(target). We considered 7 other factors(features) such as the mortgage arrears rate, real disposable income, labour force participation rate, real median household income, unemployment rate, consumer price index, and the average value of new loans.

**# Data Sources:**

We gathered our data from the CMHC website

<https://www.cmhc-schl.gc.ca/en/professionals/housing-markets-data-and-research/housing-data/data-tables/mortgage-and-debt>

https://www.cmhc-schl.gc.ca/en/professionals/housing-markets-data-and-research/housing-data/data-tables/housing-market-indicators

**# Methodology:**

We first Cleaned the Data and Analysed the relationships using a correlation matrix and other visualizations, and finally performed Machine Learning. We chose the linear Regression and Random Forest Regressor model. We also used the feature importance function of Random Forest to rank our features from most important to least. Finally, we evaluated our models on metrics such as Mean Absolute Error(MAE), Mean Squared Error (MSE), R-squared and Root Mean Squared Error (RMSE).

## **Preprocessing the data**

## Step one was to extract the data:

We had two tables one containing the Mortgage information and the other related to the housing indicators for our analysis. We decided to conduct our analysis on the country as a whole and selected 4 other provinces such as Alberta, BC, Ontario and Quebec. We created an ERD chart to check for a common link between our datasets so that they could be merged for our analysis.

## Step 2 was to clean our data:

Our preprocessing involved checking for null values in the data set, one of our tables had data in quarters and the other had annual data. So we first converted the quarterly data to an annual format and then began the rest of the steps.

Quarterly data

‘’’

‘’’

Annual data

‘’’

‘’’

Snapshot of the cleaned Mortgage data set

‘’’

‘’’

Similarly, we cleaned the housing indicators dataset. We had some columns with missing cell values. We decided to use the mean function to fill these gaps since we had a limited dataset, to begin with. This ensured that we did not lose the data in the other columns.

## Step 3 was to Transform the data

A snapshot of the correlation matrix

|  |  |
| --- | --- |
| Mortgage\_delinquency\_rate | 1.000000 |
| Residential\_Mortgage\_Arrears\_Rates | 0.964244 |
| Real\_Disposable\_Income\_Percent\_Change | 0.308583 |
| Labour\_Force\_Participation\_Rate | 0.117530 |
| Real\_Median\_Household\_Income | -0.141039 |
| Unemployment\_Rate | -0.141934 |
| Consumer\_Price\_Index\_Percent\_Change | -0.162176 |
| Years | -0.239036 |
| Average\_Value\_New\_Loans | -0.351025 |

# Correlation Matrix

‘’’

‘’’

This output is a list of correlations between the target "Mortgage\_delinquency\_rate" and the other variables/features. The correlations are expressed as coefficients between -1 and 1, where a coefficient close to 1 indicates a strong positive correlation (an increase in one variable tends to lead to an increase in the other), a coefficient close to -1 indicates a strong negative correlation (an increase in one variable tends to lead to a decrease in the other), and a coefficient close to 0 indicates little to no correlation between the variables. The values are sorted in descending order based on their absolute value.

In this case, the "Residential\_Mortgage\_Arrears\_Rates" column is found to have the strongest positive correlation with "Mortgage\_delinquency\_rate" (0.964244). The "Real\_Disposable\_Income\_Percent\_Change" column has a moderate positive correlation (0.308583), while the "Labour\_Force\_Participation\_Rate" has a weaker positive correlation (0.117530). On the other hand, "Real\_Median\_Household\_Income" has a moderate negative correlation (-0.141039), and "Unemployment\_Rate" has a similar negative correlation (-0.141934). The "Consumer\_Price\_Index\_Percent\_Change" column has a weaker negative correlation (-0.162176), while the "Years" column has a stronger negative correlation (-0.239036) and the "Average\_Value\_New\_Loans" has the strongest negative correlation with "Mortgage\_delinquency\_rate" (-0.351025).

**## Benefits of the correlation analysis:**

**### Feature Selection:** The correlations give an indication of which variables are most relevant to the target variable "Mortgage\_delinquency\_rate". For example, our goal is to build a model to predict the "Mortgage\_delinquency\_rate", the variables with high correlations can be prioritized as features in the model. This can help reduce the complexity of the model and improve its performance.

**### Identifying Multicollinearity:** High correlations between predictors can indicate the presence of multicollinearity in the data. This means that the predictors are highly correlated with each other and might be measuring the same underlying relationship. In this case, removing one of the highly correlated variables may be appropriate to reduce the model's complexity and improve its stability.

**### Understanding Relationships:** The correlations provide a general understanding of the relationships between the variables. For example, a strong positive correlation between two variables indicates that they tend to increase or decrease together, while a strong negative correlation indicates that they tend to move in opposite directions. This information can help inform the choice of predictors in a machine learning model and guide subsequent analysis.

**### Hypothesis Testing:** The correlation coefficients can be used to test hypotheses about the relationships between variables. For example, one could test the hypothesis that there is a strong positive correlation between "Residential\_Mortgage\_Arrears\_Rates" and "Mortgage\_delinquency\_rate".

# Snapshot of the scatterplot

‘’’

‘’’

The purpose of this plot is to visualize the relationship between Real Median Household Income, Consumer Price Index Percent Change, and Mortgage Delinquency Rate. By looking at the distribution of points and the size and colour of each point, we see how changes in household income and consumer prices are related to mortgage delinquency.

Based on the scatter plot, it appears that there may be some correlation between Real Median Household Income and Mortgage Delinquency Rate. As the Real Median Household Income decreases, the Mortgage Delinquency Rate tends to increase. This suggests that households with lower incomes may be more likely to experience delinquency on their mortgage payments.

The relationship between Consumer Price Index Percent Change and Mortgage Delinquency Rate is not as clear from this plot. There is a mix of points with high and low Mortgage Delinquency Rates for both low and high Consumer Price Index Percent Changes. This suggests that changes in consumer prices may not have a strong impact on mortgage delinquency.

# **Machine Learning**

**## Linear regression** We chose this since we have multiple variables as features. A snapshot of our analysis including accuracy scores and MAE, MSE, and R-Squared.

**Mean Absolute Error (MAE):** It represents the average difference between the predicted values and the actual values. A low value of MAE indicates that the model has a good fit.

**Mean Squared Error (MSE):** It represents the average squared difference between the predicted values and the actual values. A low value of MSE indicates that the model has a good fit.

**R-Squared:** It is a statistical measure of the goodness of fit of a regression model. It ranges from 0 to 1, where a value of 1 indicates that the model fits the data perfectly. A high value of R-Squared indicates that the model has a good fit.

These values give us an idea of how well our model is fitting the data. A low value of MAE, MSE and a high value of R-Squared indicate that the model has a good fit and is able to predict the target variable accurately.

Snapshot of the Linear Regression Analysis

‘’’

‘’’

## **Random Forest Regressor** we conducted this analysis to establish the importance of the features we selected.

The Root Mean Squared Error (RMSE) is the square root of the MSE. It provides an idea of the magnitude of the error, in the same units as the target variable.

In general, a higher R-squared value and lower values for MAE, MSE, and RMSE indicate a better fit of the model to the data. However, these metrics should always be considered in the context of the specific problem and data you are working with.

Snapshot of feature importance and evaluation metrics.

‘’’

‘’’

‘’’

‘’’

**# Limitations of the model**

### Small sample size – Our data set was small. If the size of the dataset used to train the model is small, it can lead to overfitting or underfitting and negatively impact the performance of the model.

### Overfitting - The model may have overfitted to the training data and may be unable to generalize well to new unseen data.

**# Future recommendations**

### Improving the factors identified as having a high impact on mortgage delinquency rates, including factors such as income, employment status, and credit score, among others.

### Continuously monitoring the market and the economy, and proactively adjusting the lending criteria and practices to keep pace with changes and minimize the risk of delinquency.

### Encouraging the collection of more comprehensive and accurate data to improve the accuracy of future predictions and decision-making.