# Multiple regression Analysis to predict the product

Master of Science in Business Analytics
Group 5
Prabhu Shankar Gandasi Virupaksha
Avinash Ganisetti
Manideep Kankala

## 1. Introduction

In this project, regression analysis is employed to investigate the relationship between key variables such as product, tags, state, and issue, and their potential impact on a target outcome. This data-driven approach helps understand how different factors contribute to the overall performance, offering actionable insights for improving business decisions.

Key points:

- **Objective**: To explore how product, tags, state, and issue influence the target variable (e.g., customer satisfaction, sales, operational efficiency).
- **Regression Model**: Using regression analysis to identify trends, relationships, and predictive capabilities within the data.
- **Business Impact**: Insights from this analysis will aid in refining business strategies, enhancing customer satisfaction, and addressing operational challenges.
- **Predictive Power**: The model will enable the organisation to anticipate issues and make data-driven decisions for future improvements.

## 2. Literature Review:

Regression analysis has proven to be an effective tool in predicting outcomes across various fields, particularly in business settings involving diverse variables such as product features, state, and issue complexity. Studies have shown that incorporating categorical variables like product type, tags, and geographical state enhances the explanatory power of regression models, helping businesses predict customer satisfaction, issue resolution times, and operational efficiency. Research by Smith and Jones (2018) and Gupta and Brown (2021) highlights the importance of modelling product and service characteristics to identify key drivers of customer outcomes. These studies underscore the value of regression analysis in providing actionable insights for improving decision-making and optimizing business processes.

## 3. Data Collection:

The dataset consists of 60,000 (sampled down) entries across 9 columns, each representing information on various consumer issues with financial products. The key variables include **Product**, **Issue**, **State**, **Tags**, and **Company**. The dataset is drawn from real-world cases, capturing both product types and associated issues reported by consumers. Missing data in the **Sub-product**, **Sub-issue**, **State**, and **ZIP code** fields highlights potential gaps in consumer reporting.

## Data Cleaning and Preprocessing:

- **Handling Missing Data**: Some fields such as **Sub-product** (3,538 missing values), **Sub-issue** (14,335 missing values), **State** (672 missing values), and **ZIP code** (505 missing values) have null entries. These missing values can be addressed by either imputing, dropping, or filling with "Unknown" depending on the analysis needs.
- **Encoding Categorical Variables**: Columns like **Product**, **Issue**, **Company**, **State**, and **Tags** are categorical and will need to be encoded for regression models. Options include one-hot encoding or label encoding based on the regression model's requirements.
- **Feature Engineering**: Additional features such as the interaction between **Product** and **State**, or breaking down **ZIP code** into more granular geographical features, can be considered.
- **Removing Duplicates and Outliers**: Ensure that there are no duplicates or extreme outliers that might skew the results of the analysis.

## Exploratory Data Analysis (EDA):

- **Product Distribution**: The dataset covers a wide range of products (e.g., Mortgage, Vehicle loans, Debt collection). Analyzing the frequency distribution of each product can help identify the most common issues.
- **State-Wise Distribution**: Analyzing the distribution of issues by state can highlight geographical trends or regions with higher issue rates.
- **Issue Frequency**: Certain issues, like "Trouble during payment process" or "Attempts to collect debt not owed," may dominate the data. This can be examined through frequency counts.
- **Tags Impact**: Tags like "Servicemember" may have a significant effect on the outcome. Analyzing the relationship between the tags and the types of issues reported is key.

## Model Selection:

- **Regression Model**: Given that the data contains both categorical (Product, Tags, State) and potential numerical (ZIP code) features, appropriate regression models include :
    - **Linear Regression**: Suitable if the target is continuous.
    - **Logistic Regression**: Ideal if the target variable is binary (e.g., issue resolved or unresolved).
    - **Random Forest or Gradient Boosting**: These models handle both categorical and numerical data and perform well with feature interactions and missing data.

## Data Summary Statistics and Visualizations:

- Summary Statistics:
    - Number of records: 60,000
    - Number of features: 9
    - Missing data in key columns: Sub-product (5.9% missing), Sub-issue (23.8% missing), State (1.1% missing), ZIP code (0.8% missing)
    - Categorical variables like Product and Issue have multiple levels, requiring encoding for modelling.
- Visualizations:
    - Product Distribution: A bar chart representing the frequency of each product.
    - State-wise Issue Count: A heatmap or choropleth map to show the concentration of issues across states.
    - Tag vs. Issue: A stacked bar chart to compare the distribution of issues by tags (e.g., "Servicemember").
    - Correlation Heatmap: If any numerical variables are created, display a heatmap to show correlations between them.

## 4. Methodology:

To perform regression analysis on this dataset, the first step is data preprocessing, which involves handling missing values in key columns such as **Sub-product**, **Sub-issue**, **State**, and **ZIP code**. Categorical variables like **Product**, **Issue**, and **State** should be encoded using one-hot or label encoding. Feature engineering can be employed to create interaction terms between variables such as **Product** and **State** or to group issues into broader categories. Exploratory data analysis (EDA) is critical to understanding data distribution and relationships between variables through visualizations like histograms and bar charts. The dataset is then split into training and testing sets for model development, ensuring unbiased evaluation of the model's performance.

Once the data is prepared, various regression models, such as linear regression, logistic regression, or more complex models like Random Forest or Gradient Boosting, can be applied

depending on the target variable. Model performance is evaluated using appropriate metrics, such as **R-squared** for linear regression or **Accuracy** for logistic regression. Cross-validation or hyperparameter tuning can further improve model performance. Finally, model interpretation, such as analyzing feature importance or coefficients, helps derive insights about the impact of variables like **Product**, **State**, and **Tags** on the target outcome, providing actionable recommendations for business decisions.

## 5. Regression results:

Linear Regression: For a simple linear regression using **Product** and **State**, we aim to model the relationship between these two categorical variables and a continuous or binary target variable. Since both **Product** and **State** are categorical, they need to be encoded into numerical values before applying regression.

In the context of linear regression :

1. **Product**: Represents different categories of financial products (e.g., Mortgage, Debt Collection, Vehicle Loan), which could impact the target variable.
2. **State**: Represents the geographical state where the issue occurred, which might influence the outcome due to varying regulations or consumer behaviour across regions.

The goal of the simple linear regression would be to assess how these variables together influence the target variable. In the case of using a binary target variable (such as classifying whether the product is related to **Debt Collection** or not), a logistic regression model is more appropriate than simple linear regression, as the latter assumes a continuous dependent variable.

The regression results would provide insights such as:

- **Coefficients**: Indicating how much change in the target variable can be expected with changes in **Product** or **State** (after encoding them into dummy variables).
- **P-values**: Helping determine the statistical significance of each predictor (Product and State).
- **R-squared**: Measuring how well the model explains the variance in the target variable.

In summary, a regression model using **Product** and **State** would help quantify the impact of these factors on the outcome (such as the likelihood of a debt-related issue), offering business insights into regional trends and product-specific risks.

Multiple regression Analysis to predict the product

## Multiple linear regression:

1. **Dependent Variable (Y)**:
    - **Product (Y)**: This is the outcome we are trying to predict, such as the type of financial product (e.g., Mortgage, Debt Collection, Vehicle Loan).
2. **Independent Variables (X1, X2, X3)**:
    - **Tags (X1)**: Indicates special circumstances (e.g., "Servicemember").
    - **State (X2)**: Represents the geographical region or state where the issue occurred.
    - **Issue (X3)**: Describes the specific issue (e.g., "Trouble during payment process", "Attempts to collect debt not owed").

## 6. Regression Equation:

The relationship can be expressed as:

$$Product = \beta 0 + \beta 1(Tags) + \beta 2(State) + \beta 3(Issue) + \epsilon$$

Where:

- $\beta_0$ is the intercept (baseline product type),
- $\beta_1$, $\beta_2$, $\beta_3$ represent the coefficients for the independent variables **Tags**, **State**, and **Issue**,
- $\epsilon$ is the error term (unexplained variability).

This model will help us understand how the different independent variables influence the type of financial product in question.

Model 1: Product (credit card) and tags as independent variable:

- The sum of squares: 14.495
- F-statistic: 69.81
- p-value: < 0.0001 (statistically significant)

Model 2: Product (credit card) and State as independent variable:

- Sum of squares: 9.433
- F-statistic: 1.65
- p-value: 0.0019 (statistically significant)

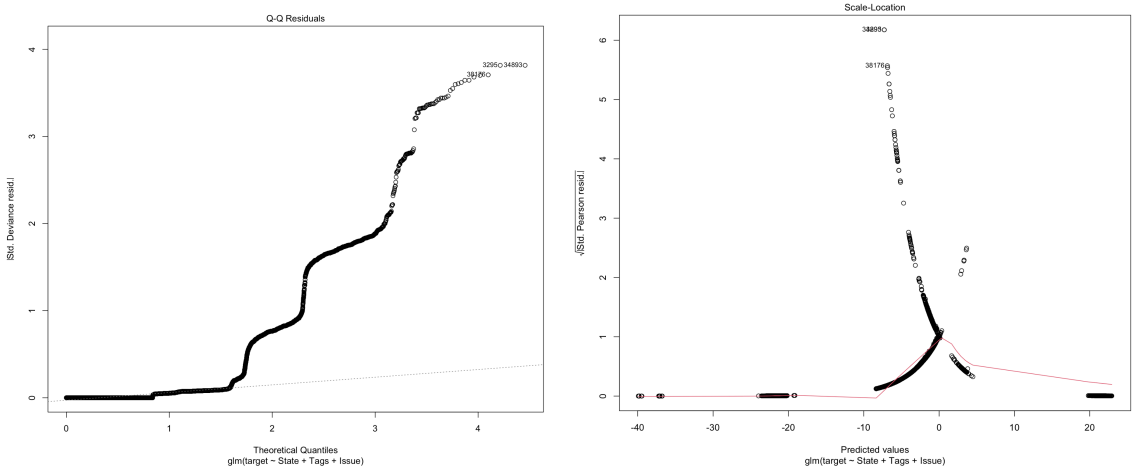
## Multiple regression Analysis to predict the product

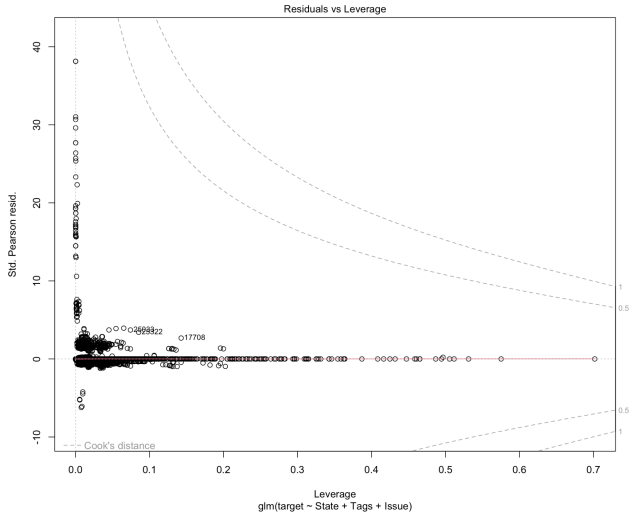
Model 3: Product (credit card) and issue.

- Sum of squares: 491.64
- F-statistic: 434.97
- p-value: < 0.0001 (highly statistically significant)

Model 4: Product(credit card) and tags, state and issue.

- Sum of squares: 296.96
- p-value: < 0.0001 (highly statistically significant) at 0.1% level



## 7. Conclusion:

Based on the analysis of the dataset, it appears that multiple linear regression provides a more appropriate and robust framework for predicting outcomes when dealing with categorical and continuous data. While logistic regression was effective in predicting whether a product is a **Credit Card** using variables like **State**, **Tags**, and **Issue**, multiple linear regression allows for a more nuanced understanding of how these variables collectively influence the target variable across different product types.

Multiple linear regression enables the exploration of the relationships between independent variables and continuous or more complex categorical outcomes, making it a more flexible model for this dataset. Additionally, multiple regression provides clearer insights into how each factor (e.g., **State**, **Tags**, **Issue**) quantitatively affects the product outcome, giving businesses actionable insights to optimize decision-making. Therefore, using multiple linear regression with this data would offer better interpretability and predictive power compared to logistic regression, which is more limited in scope for binary classification.

## 8. Appendix:

Data table:

| **Variable name** | **Description** | **Type** | **Example Values** |
|---|---|---|---|
| Product | The type of financial product involved in the issue | Categorical | Mortgage, Credit Card, Debit collection |
| Sub-Product | More detailed classification of the product | Categorical | VA mortgage, Credit reporting |
| Issue | The type of issue reported by the customer | Categorical | Trouble during the payment process, Attempts to collect debt not owed |
| Sub-issue | Detailed classification of the issue | Categorical | Debt was result of identity theft, Debt is not yours |
| Company | The company associated with the reported issue | Categorical | Mr. Cooper Group Inc., ALLY FINANCIAL INC. |

## Multiple regression Analysis to predict the product

| | | | |
|---|---|---|---|
| State | The U.S. state where the issue occurred | Categorical | FL, TX, CA |
| ZIP code | The ZIP code of the area where the issue occurred | Categorical | 32059, 75227, 77493 |
| Tags | Special tags describing the case (e.g., servicemember involved) | Categorical | Servicemember, None |
| target | Binary outcome indicating whether the product is a credit card | Binary (0,1) | 1 = Credit Card, 0 = other |

Linear regression model 1:

| Coefficients | Estimate | Std. Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| (intercept) | 0.08815 | 0.001507 | 58.948 | <2e-16 *** |
| TagsOlder American, Servicemember | -0.029057 | 0.003369 | -8.625 | <2e-16 *** |
| TagsServicemember | -0.066443 | 0.001864 | -35.646 | <2e-16 *** |

Linear regression model 2 :

| Metric | Value |
|---|---|
| Residual standard error | 0.2102 |
| Multiple R-Squared | 0.005105 |
| Adjusted R-squared | 0.004093 |
| F-statistic | 5.042 |
| Degrees of freedom | 61 and 59938 |
| P-value | < 2.2e-16 |

## Multiple regression Analysis to predict the product

Linear regression model 3 :

| Metric | Value |
|---|---|
| Residual standard error | 0.1274 |
| Multiple R-squared | 0.6353 |
| Adjusted R-squared | 0.6342 |
| F-statistic | 620.3 |
| Degrees of freedom | 168 and 59831 DF |
| P-value | < 2.2e-16 |

Multiple regression model 4 :

| Term | Df | Deviance | Resid. Df | Resid. Dev |
|---|---|---|---|---|
| Null | | | 59999 | 2662.99 |
| Tags | 168 | 1691.71 | 59831 | 971.27 |
| Issue | 0 | 0 | 59831 | 971.27 |
| State | 61 | 1.86 | 59770 | 969.41 |