Student Declaration:

By submitting this assignment, I confirm that I am aware of DBS's policy regarding cheating, plagiarism and all other forms of academic impropriety. The coursework submitted is my own or my group's work, and all other sources consulted have been appropriately acknowledged. I am aware that in the case of doubt an investigation will be held.

Student Name: Prabin Chaudhary **Student No.: 20032753**

Assignment Title: Data Visualisation for Loan Eligibility Prediction Dataset

Date of Submission: 29th July 2024

Introduction

In today's world, there is abundance of data and the same data can be used to predict if a customer is good or bad and eligible or not which is effective in managing risk of losing money by financial institutions due to defaults. For making these decisions, various data analytics tools are available and data visualization can play a crucial role in decision making. It is a base for good data storytelling and finding the various trends and anomalies present in today's financial transactions.

The Dataset

For the purpose of data visualisation, an interesting dataset related to loan eligibility prediction has been chosen from Kaggle. The link for the dataset is: https://www.kaggle.com/datasets/shadabhussain/credit-risk-loan-eliginility. The dataset has 45 columns/features and 63,999 rows or data points. The dataset provides detailed information about the demographics of customers namely employment title and length, home ownership, annual income and others, financial position, outcome of the loan applications and others. Therefore, this dataset can be used for various analyses such as the effect of annual income or homeownership on loan approval, risk and trend associated with loan approval, loan status in various states of The USA and many others.

Main User (Who it's for)

The main users of the visualization are loan officers and Credit Risk Analysts which are explained as follows:

1. **Loan officers**: The dataset is used by loan officers for the efficient evaluation and approval of loan applications. They assess the creditworthiness of borrowers by examining factors such as employment length, annual income, home ownership status, and debt-to-income ratios. Identifying patterns in borrower characteristics and understanding the determinants of loan approvals or rejections enables them to customize lending criteria in order to improve approval rates and reduce defaults. With the use of this dataset and visualization, the loan officers can make correct decisions and make sure the loan is given to right candidates while reducing the risk of defaults.

2. **Credit Risk Analysts:** Being the head of credit department, they are responsible for analyzing and determining if the credit department is at risk or not. They are involved in policy and strategies making and credit scoring as well. So, the dataset and visualization can be an important tool in the analyses of the credit customers and save the company from suffering huge financial losses resulting from default of loan.

Data Operations

Data transformations are necessary to get the accurate results in any dataset. The dataset was mostly clean. However, the 'Addr State' column was set the Geographic role of State/Province which gave a proper result while plotting the map. Similarly, three new columns were created namely 'Loan Approved', 'Loan Rejected' and 'Loan Status Count' from column 'Loan Status' which were used to show the loan approved and rejected based on verification status, purpose, states and home ownership.

Justifications

The dataset was selected for its comprehensive coverage of loan applications, including detailed borrower information, loan amounts, purposes, verification statuses, and outcomes (approved or rejected). This detail enables detailed analysis of factors influencing loan approval and risk assessment, essential for financial institutions.

The reason to choose Tableau was due to its user-friendly drag-and-drop interface, which allows for the creation of complex visualizations without extensive programming knowledge. Tableau's wide categories of visualization options and customization features enable users to develop highly informative and visually appealing dashboards. Additionally, another advantage of Tableau is that it is connected to large data sources and has the ability to handle large datasets as well. Tableau is also popular for its dashboards which are interactive and allow users to view and explore the data more clearly in a responsive way.

Among 45 features, for the purpose of the visualization, 5 columns were chosen namely Loan Amount, States, Purpose, Verification Status, Loan status. Two new columns named Loan Approval and Loan Rejected were formed from Loan Status which have been used for the data visualization. The reasons behind selection of these columns are to determine the following:

1. Relation between loan amount and loan status in various states of the USA
2. For which purpose there is more demand of loan?
3. Which loans are being rejected the most on the basis of purpose?
4. Which state has the most loan rejections?
5. What effects do verification status and home ownership have in determining the loan approval rate?

Among the 6 visualisation charts, the most interesting chart seems to be the chart showing the relationship between loan amount and states. This chart represents the total loan demand in different states and which has the lowest demand. The states with the low loan amounts can be marked as new hub for promotion campaigns as these places are still uncharted showing the business opportunities.

Insights

1. Looking at the charts, California seems to have the highest demand of loan as well as the loan rejection as well. Similarly, even though the area of Alaska is the biggest among all the states, the loan demand seems to be low and it is about 1.73% of the loan demand of California which is much smaller in area than that of Alaska. However, percentage wise the loan approval rate of Alaska is somewhat similar to that of California which is approximately 29% for Alaska versus 27% for California.
2. Similarly, the highest demand of loan is for the purpose of debt consolidation as per the visualization. That is to say, there is high loan demand to repay their previous loans which does not seem to be good. The loan for debt consolidation is approximately 3.52 times the second highest loan demand which is for credit card loans. The chart clearly shows that there is high demand of loan in unproductive sectors such as car, credit card, home improvement and others than productive sectors like small businesses, medical and others.