DV Assignment 3: A Visual Analytics Workflow to Understand Financial Risk

Prateek Rath
IMT2022017
IIIT Bangalore
prateek.rath@iiitb.ac.in

Mohit Naik
IMT2022076
IIIT Bangalore
mohit.naik@iiitb.ac.in

Anurag Ramaswamy
IMT2022103
IIIT Bangalore
anurag.ramaswamy@iiitb.ac.in

Introduction

Through this assignment, we aim to gain a deeper understanding of Financial Risk and the factors affecting it, by implementing three separate visual analytics workflows, based on the Visual Analytics Workflow proposed by Keim et al. [1]. Each workflow consists of a number of visualizations to visualize the data, an instance of a machine learning model to make predictions on the data or cluster the data, followed by insights and knowledge gained from the combination of the visualizations and the model, and a feedback loop which executes a suitable data transformation.

The dataset we used is the same as the one used in Assignment-1 [2], a kaggle dataset which contains financial data of a large number of users [3]. We used the same subset of 5000 entries which we used in the first assignment. To supplement that, we added a part of another similar dataset [4]. The only preprocessing done was to have identical columns in both datasets and join them. We thus get a combined dataset with 10000 rows and 28 columns.

Workflow 1 - Analysing Model performance and behaviour on Loan Approval predictions

Introduction

Machine learning is starting to play an increasing role in predicting loan approval outcomes. Although it is not commonly used as a direct stand-alone method, it continues to be of importance in serving as a reference. Here, we analyze model behaviour on our Loan Approval dataset. Using the visual analytics loop, we aim to improve model performance, assess its limitations, and ultimately develop a model that can reliably determine loan approval.

A. Notes for the reader

The images in this workflow may appear blurry. The reason for this is that most are screen clippings taken from actual plot images that contain a large number of subplots. We wish to focus on single subplots here to see things better.
All the actual images can be found on github or the one-drive link.
Since this workflow has different images in the report from that of the original repository, the numbering notation might be slightly different. This however will not lead to any breach in understanding.

The Workflow

The workflow is shown below in figure 1.

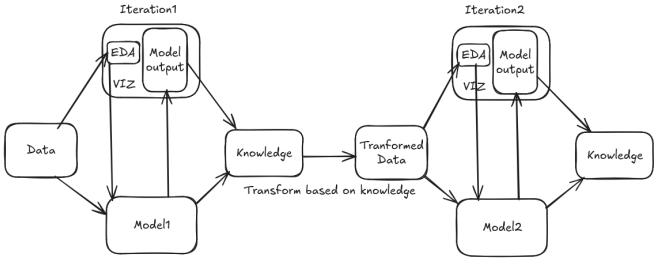

Fig. 1. Basic idea of the workflow

The workflow consists of the following steps:

1) Preprocessing the data and getting it ready to work on after merging the two datasets.
2) Exploring the data using EDA, done in assignment-1 and here too
3) Choosing and Training a model based on EDA
4) Visualizing the model output and diagnosing failure.
5) Gaining knowledge from the model output visualizations.
6) Transforming the data via feature engineering(via common sense and model insights)
7) Performing EDA on new features.
8) Training a second model
9) Visualizing the model output and diagnosing failure.
10) Gaining knowledge from the model output visualizations.
11) Repeat the above any number of times, we stop at two iterations.

EDA1

To understand the nature of the dataset, we make use of proportion stacked bar charts and GPLOMs(Generalized Plot Matrices). Since, the data consists of both categorical and continuous columns, GPLOMs seemed appropriate to do the

EDA. We didn't have the chance to use them in assignment-1(we hadn't learnt about them) and hence chose the option here.

GPLOMs: Figure 2 shows an image of the GPLOM.

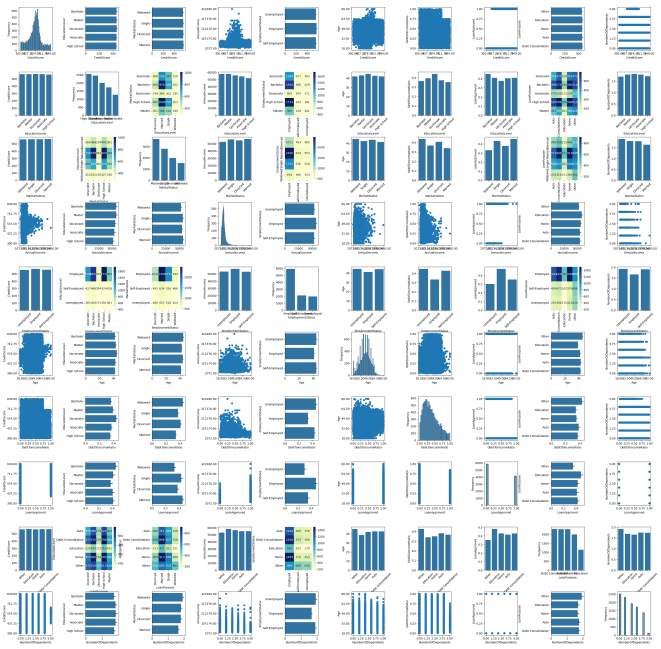

Fig. 2. GPLOM

Of course we haven't plotted all the features in the GPLOM, that would be too many as the number of plots in the gplom in $O(n^2)$ where n is the number of columns. We observe the following from the GPLOM. Most variables have no correlation with others. That is we see for example that credit score has no relation with education level. The mean credit score is the same in each category and is independent of the level of education. The frequency distributions(seen along the diagonal) seem to be normal distributions with different means and distributions(skewed). For instance 'DebtToIncomeRatio' has a skewed normal distritbution. All these things influence our choice of the model. Further details on the choice of the first model are provided in the sections that follow.

Stacked Proportion Charts: The fancy term 'Stacked Proportion Charts' just refers to stacked bar charts and histograms. Each bar is of the same height. The proportion of the bar colored red indicates the proportion of people in that category that were denied loans and the rest of the portion is colored green to indicate loan approval. Two plots for the same are shown in Figure 3 and Figure 4.

Note that the ranges of the bins are not important, rather notice the trends. We observe the following from the stacked proportion charts all of which can be found in the notebook on github.

The visualizations above via stack bar charts show some clear trends:

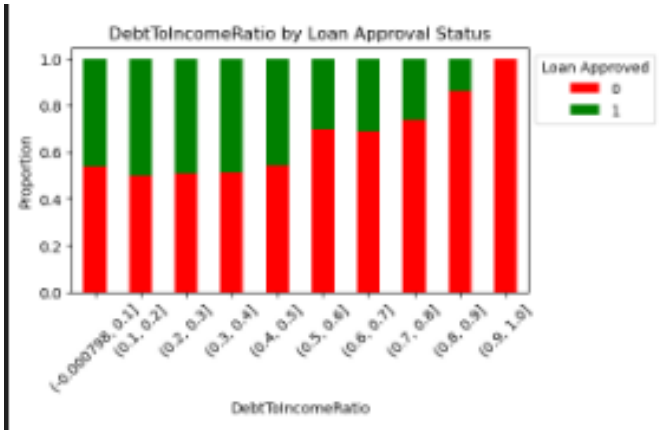

Fig. 3. Stacked Proportion Chart for DTI Ratio

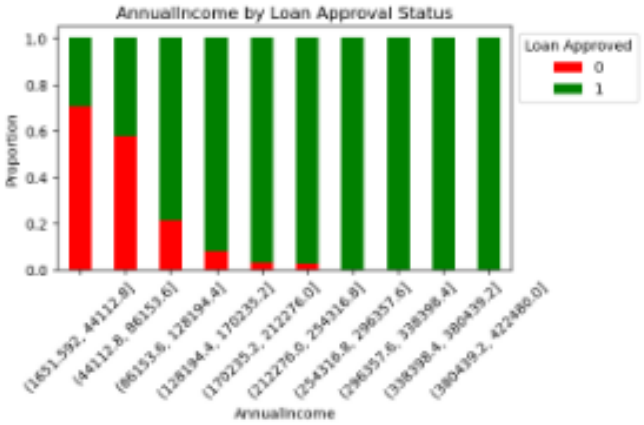

Fig. 4. Stacked Proportion Chart for Annual Income

1) If annual income is high enough, people have no problem getting loans.
2) Low loan amounts are usually approved.
3) High DTI ratios are usually rejected.
4) Higher credit card utilization usually leads to rejection.
5) A higher number of open credit lines leads to approval.
6) A previous loan default is likely to result in rejection.
7) A very high savings account balance leads to loan rejection.
8) A high net worth leads to loan approval.

Most other factors show no clear trend.

Model1: GNB

For our first model, we choose Gaussian Naive Bayes. This choice comes from our observations of independence and normal distributions in the GPLOM. We can also choose to drop features that aren't of high importance in influencing the 'LoanApproved' target variable. However, we avoid doing this. Apart from decreased model accuracy the other reason for this is that our naive assumption already is very simplifying. Further simplifying assumptions that some features don't matter at all would be unreasonable. We train the model on the data using the typical process of creating a validation set

from a subset of the actually available training data. However, we end up getting extremely poor accuracy scores of just over 68 percent on both test and validation data.

Diagnosis based visualizations

In this part of our workflow, we try to understand through visualizations, where our model is performing incorrectly. Specifically we make use of two visualizations horizontal feature distribution box plots and parallel coordinate plots.

1) Horizontal Box Plots: Figure 5 and Figure 6 show such boz plots.

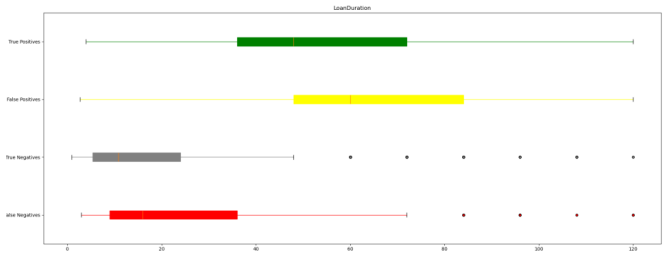

Fig. 5. Horizontal Box Plots [5] for Loan Duration

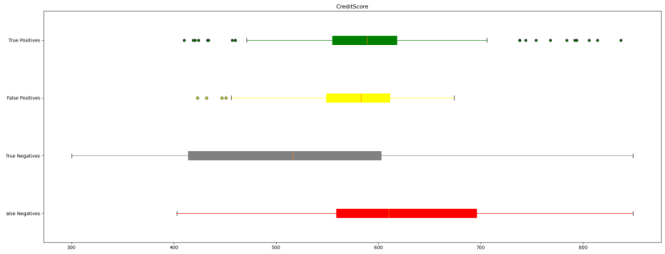

Fig. 6. Horizontal Box Plots for Credit Score

The idea of horizontal box plots to visualize feature distributions is not new. It is a neat technique to diagnose errors in classification models [5]. Each figure actually has four box plots within it, each showing the range of the feature for different kinds of data points. The greens indicate true positives, the yellows indicate false positives, the greys indicate true negatives and the reds indicate false negatives. The position of the bar shows the range of the feature for which a particular class(TPs, TNs, FPs or FNs) is prominent. Notice that in Figure 5 the positives are very far away from the negatives in the feature range. This shows that Loan Duration is a good metric when used alone to classify individuals. On the other hand if we take a look at credit score, to our surprise, we see a lot of false negatives at higher Credit Scores. This might be because our naive model implicitly places higher weight on LoanDuration or other columns and denies these individuals with high credit scores loans that they deserve. Maybe these individuals have low annual incomes or are asking for very large loan amounts. Clearly, our naive bayesian assumption fails to capture these trends.

Brushed Parallel Coordinate Plots: Here we plot parallel coordinate plots for a number of columns. We focus only on the true positives and false positives in this report, a similar analysis can be done by seeing the plot on false negatives and true negatives. Figure 7 shows our brushed parallel coordinates plot.

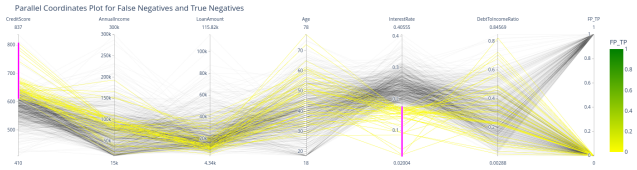

Fig. 7. Brushed pcp showing TPs and FPs

We have brushed regions of low interest and high credit score as these correspond to ranges where the false positives are concentrated without any overlap with the true positives. There seem to be quite a few lines on which the model is going wrong, even though they have high credit score. One notable reason could be their low annual incomes, but we can't say for sure as we haven't plotted all the columns in the pcp. However, this again tells us to reconsider our naive assumption.

Overall through this diagnosis phase we realize that our naive assumption may not be ideal for the following reasons:

1) Inherent randomness in the dataset
2) No consideration of a combination of features: In real life, a loan is never given based one metric or a few metrics. There are thorough background checks, and people even look at features such as marital status, number of dependents and even age. Our naive bayes assumption may take these considerations a bit too lightly.
3) In nature most distributions are gaussian. However, this alone cannot push us to expect excellent results from gaussian models.

The Transformation

We have now realized that our model is not well suited for the given dataset. It produces a low accuracy score and makes mistakes where it shouldn't due to its simplifying assumptions. Hence we have to choose a new model for our next iteration. Since we don't need our independence assumption any more, we can create engineer new features that combine previous ones. We generate the following additional features:

1) LoanAmountPerIncome
2) LoanDurationToAge
3) LoanAmountToCreditScore
4) NetWorth: Assests - Liabitilites
5) CredRatio: NumberOfCreditInquiries / LengthOfCreditHistory
6) Familia: NumberOfDependents + (Marital Status == 'Married')

7) AnnualIncome_mul_PaymentHistory
8) CreditScore_and_DTI

Some of these new features seem fairly intuitive and obvious while others aren't. For instance "CredRatio" looks like a neat option as it normalizes the number of inquiries based on the length of credit history.

EDA2

We don't do much in this section except visualizing the proportion stacked charts for the newly engineered variables. We show one of these in Figure 8.

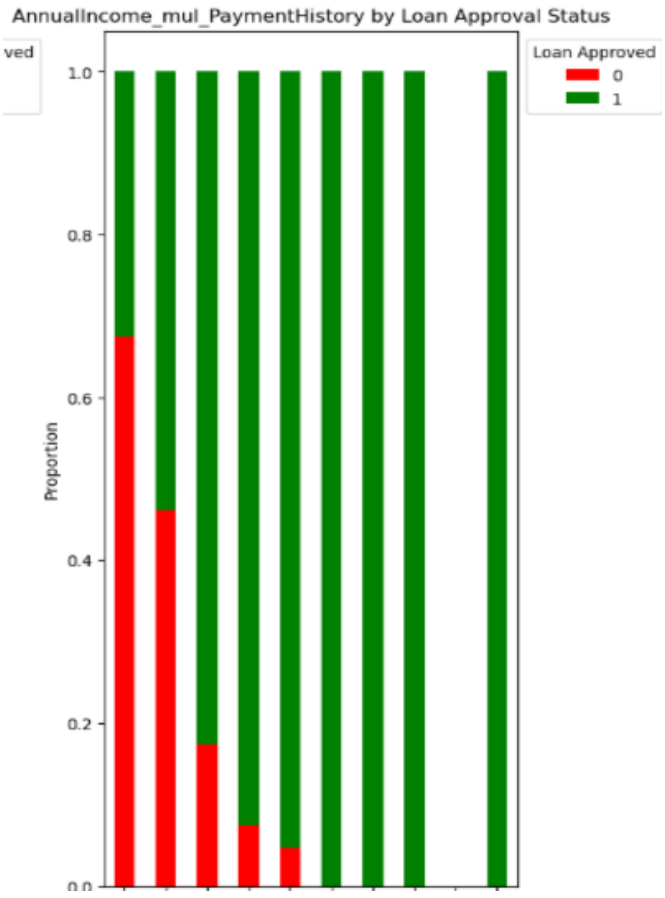

Fig. 8. Proportion stacked chart for AnnualIncome_mul_PaymentHistory

Model2: XGB

The model we choose here is the extreme gradient boost. Considering the inherent randomness in our dataset and the fact that we need to consider every feature, a model that improves from its mistakes seems like a great attempt to improve accuracy. Indeed this is the case and we get an accuracy of greater than 95 percent on training data and about 83 percent on validation data.

Visualizing the new model's predictions

Once again we plot stacked bar charts and try to compare our current model's predictions as opposed to the previous one. This time we see that in Figure 9, the credit score no longer has false negatives at higher values. Somehow this model overcomes the shortcomings of the previous ones and makes better predictions. The other box plots are present in the repository and can be viewed from there.

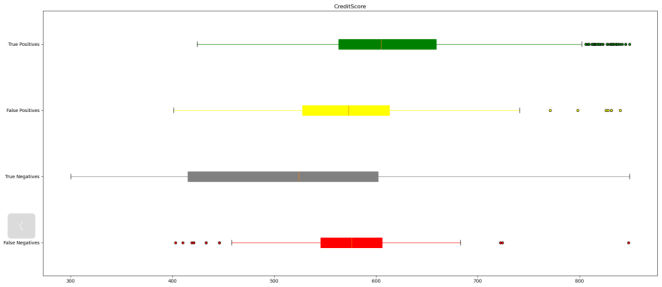

Fig. 9. Improved credit score horizontal box plot

We also used shapley to visualize some other plots such as feature importance. Figure 10 shows this as a bar chart.

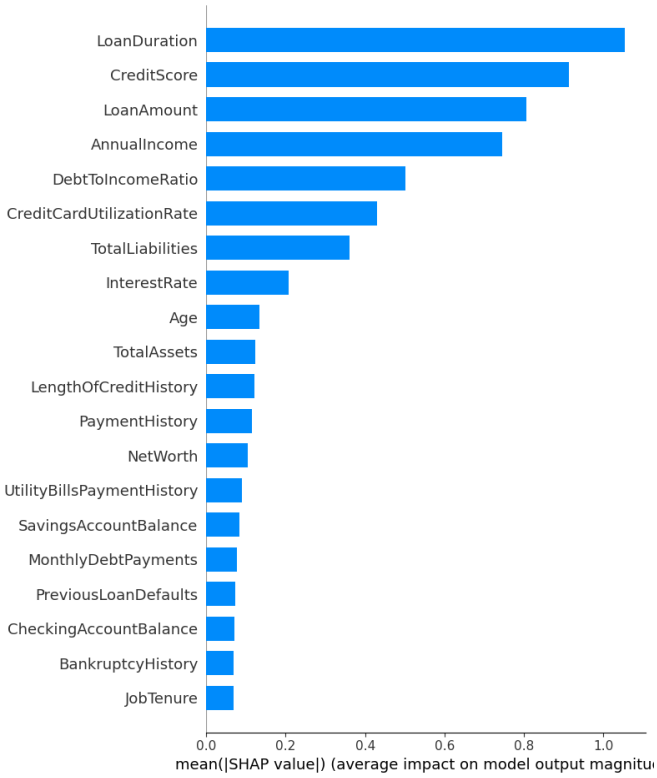

Fig. 10. Feature Importance

We see that Loan Duration seems to be the most important feature. This could be true, but this could also be an over-fitting on this dataset. In the next section we discuss potential

remedies for future iterations.

To see how this model performed, we decided to try and run it on some 'good', easy to judge data and then on some 'ambiguous' or not so easy to judge data. The good data had features that usually had a clear outcome, they had individuals with low DTI, high credit scores, high annual income etc. Some other features were sampled at random. The model performed extremely well on such good clear data points that we generated. On the other hand on ambiguous points where it was relatively harder to figure out whether the loan was to be given or not, the model still performed with an accuracy of over 65 percent.

Conclusion for workflow 1

Our model performs well in most cases. With an accuracy of 83 percent on a partly randomized dataset, we are doing pretty well. Here we suggest some improvements for further iterations.
Firstly, loan duration is the feature that is considered highly important. We may not want this to hold while using a transparent model in a real life scenario. Potential options include dropping the feature all together or coming up with transformations that somehow provide low weight to this feature. Next we talk about post classification rules. Post classification rules are rules that are applied over a model's predicted output to get the final output. For instance a rule might state that if someone has a very high annual income, we need to give him/her a loan irrespective of the fact that they have defaulted on loans before. Such rules can arise from two scenarios. The first is bank or lender policy or government laws. The second way is a little more sophisticated. It involves looking at the horizontal box plots and finding ranges in which only the false negatives or false positives lie. Then we can forcefully turn the outcome to the desired value for such ranges where the model is a little confused.

Workflow 2 - Analysing Model performance and behaviour on Credit Score predictions

Credit scoring systems serve as the backbone of modern lending decisions, playing a pivotal role in determining an individual's creditworthiness and financial opportunities. These scores, which reflect a person's credit history and financial status, have far-reaching implications beyond just loan approvals - they influence interest rates, insurance premiums, rental applications, and even employment opportunities in some sectors. The ability to accurately predict credit score categories is therefore crucial for both financial institutions and consumers. For lenders, precise credit score category predictions help minimize risk exposure and optimize lending strategies, while for consumers, understanding how their financial decisions impact their predicted credit score category can lead to better financial health.

To analyze model performance in credit score category prediction, we follow a structured approach that begins with comprehensive data analysis, moves through visualization and knowledge extraction, implements initial modeling, performs data transformation based on insights gained, and concludes with model refinement. This iterative workflow, illustrated in Fig.11, aims to not just predict credit score categories but to understand the underlying patterns that drive these predictions. Focusing on credit score categories - low ($<$ 500), medium ($>=$ 500 & $<$ 700), and high ($>=$ 700) rather than exact scores provides more actionable insights, as most lending decisions and financial products are structured around these broader credit score bands rather than precise numerical values. I have performed 2 iterations of this workflow.

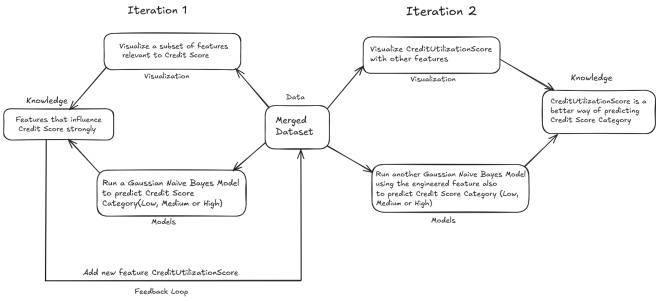

Fig. 11. Diagram illustrating the workflow

To start iteration 1 of the workflow, I visualize credit score and the factors related to it in a variety of ways. First, I visualize the distribution of credit score categories in the dataset via a bar chart as shown in Fig.12. This shows that majority of the people have a moderate credit score followed by a lesser number of people having low credit scores and extremely few people having a high credit score.

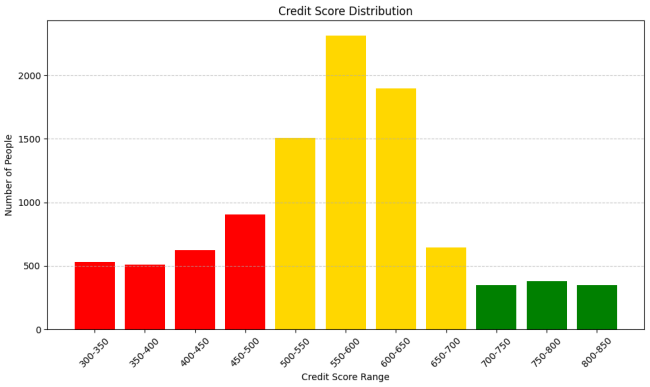

Fig. 12. Bar chart showing the distribution of credit scores

Based on the line chart showing the distribution of credit scores by loan approval status (denoted by 0 and 1) as seen in Fig.13, we can observe that the data reveals distinct patterns between approved and denied loans. The majority of approved loans (orange line, marked as 1) show a strong peak around credit scores of 600, with a right-skewed distribution. In contrast, denied loans (blue line, marked as 0) display a more spread-out distribution with a lower peak also around 600,

but with notably more density in the lower credit score ranges (300-500). This visualization suggests that higher credit scores are generally associated with loan approval and vice versa.

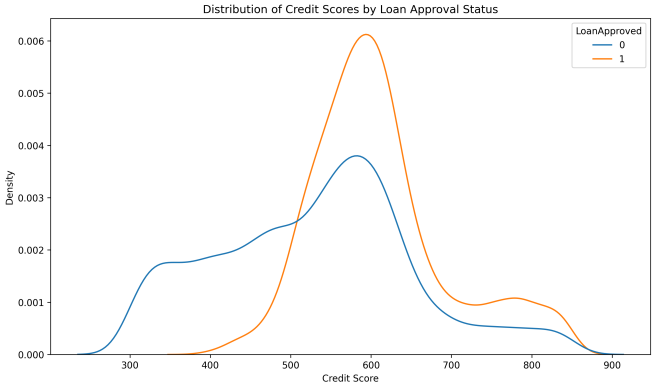

Fig. 13. Line chart showing the distribution of credit scores for approved and denied loans

In Fig.14 we can see that for a lower credit score (< 500) almost all the loans are rejected irrespective of their debt-to-income ration and the loan approvals increase as we move towards areas showing high credit score. Notably, we can see that there are no people in the top right corner of the graph, indicating that there are no people with both a credit score higher than 700 and a debt-to-income ratio higher than 0.7. This visualization further confirms our observation about the relation between credit scores and loan approvals as seen in Fig.13 and improves our knowledge on the relationship between debt-to-income ratio and credit scores.

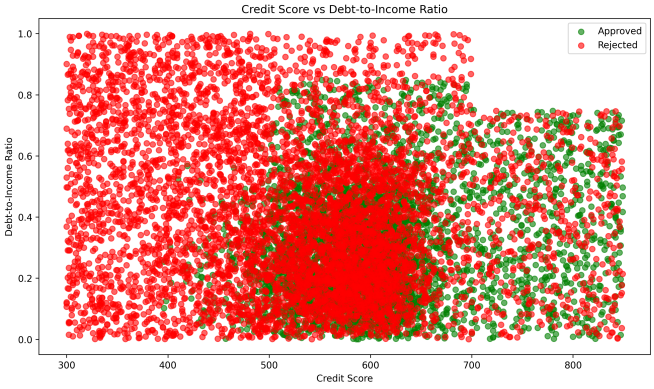

Fig. 14. Scatter plot showing credit score by debt-to-income ratio, colored by loan approval

A credit inquiry is a request for credit report information from a credit bureau. In general, there is an inverse relationship between credit score and the number of credit inquiries. Individuals with a higher credit score tend to have fewer credit inquiries, as they are seen as lower-risk borrowers by lenders. Frequent credit applications and inquiries can have a negative impact on a person's credit score, as it may indicate financial instability or a higher risk of defaulting on loan payments. Fig.15 and Fig.16 provide offer further insights into this relationship. Fig.15 shows that as job tenure (the number of years a person has been employed) increases, the total number of credit inquiries decreases. This suggests that individuals with more job stability and longer employment histories tend to have fewer credit inquiries, likely because they are seen as lower-risk borrowers by lenders. Fig.16 reveals that middle-aged individuals, especially those with low or medium credit scores, tend to have the highest number of credit inquiries, potentially due to major financial burdens and factors like more active borrowing or financial instability during this life stage. Consistent with the general trend, individuals with high credit scores consistently have fewer credit inquiries compared to those with low credit scores, across all age groups.

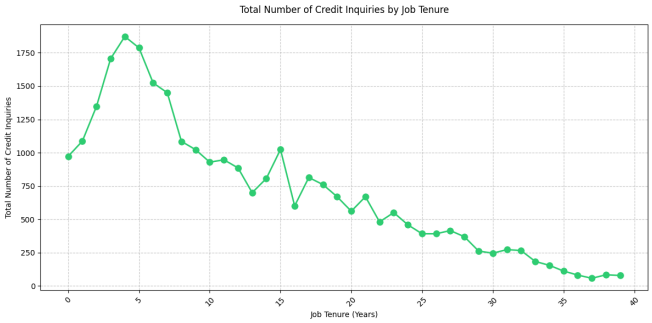

Fig. 15. Line chart showing Number of Credit Inquiries vs Job Tenure

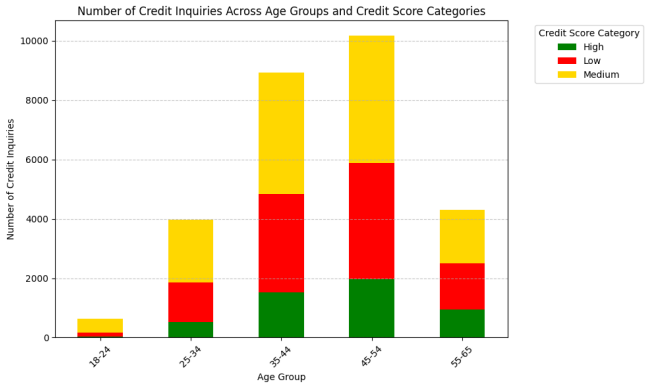

Fig. 16. Stacked bar chart [6] showing number of credit inquiries across different age groups coloured by credit score category

To look at the impact of another attribute, the number of open credit lines, a parallel coordinates plot was plotted as seen in Fig.17 and I couldn't find any relationship between this attribute and the previously mentioned attributes(Loan Approved, Number of Credit Inquiries and Debt-To-Income ratio). This might be due to the fact that the data is synthetically generated or that it actually does not have any correlation with the aforementioned attributes. In any case, I have not used this attribute for the data transformation step. Our observation that people with high debt-to-income ratios (> 0.7) do not have high credit scores ($>= 700$) as seen in Fig.14 is further justified when we look at the brushed image of the parallel coordinates plot as seen in 18.

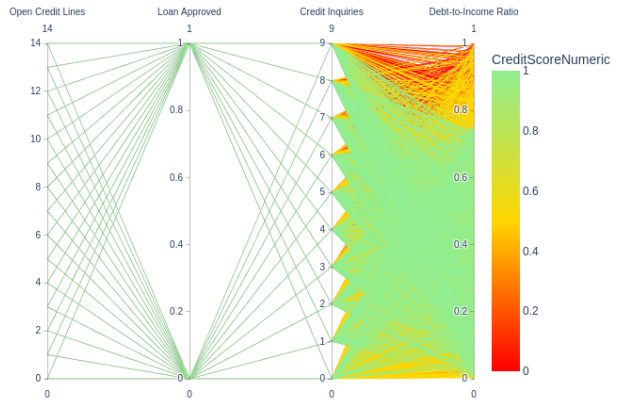

Fig. 17. Parallel coordinates plot to explore the relationships of Number of Open Credit Lines with other factors related to Credit Score

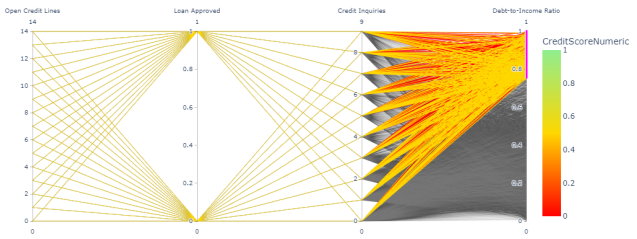

Fig. 18. Brushed version of parallel coordinates plot to explore the relationships of Number of Open Credit Lines with other factors related to Credit Score

Now, we are done with the visualization step of iteration 1. We now run a Gaussian Naive Bayes model on a subset of the relevant features which was decided from the knowledge gained in the above visualization step. These features include 'Age', 'JobTenure', 'NumberOfCreditInquiries', 'LoanApproved' and 'DebtToIncomeRatio'. The overall accuracy of the model came out to be 68.2%.

Now, we add a feedback loop i.e. data transformation based on our knowledge as follows -

Add a feature called *CreditUtilizationScore*, which is a weighted sum of the following:

- *NumberOfCreditInquiries*
- *DebtToIncomeRatio*
- *LoanApproved*
- *Age*

It assigns negative weights to Age, NumberOfCreditInquiries and DebtToIncomeRatio while assigning a positive weight to LoanApproved. I have not added Job Tenure here as I felt that since I want to show an increase in accuracy using the Gaussian Naive Bayes model after data transformation, it would not be ideal to take both JobTenure and Age since they both are highly correlated and this model assumes all features are inherently independent.

Now in iteration 2, we visualize the overall relationships of the new feature with its components and colored by credit score in the scatter plot matrix as shown in Fig.19. CreditUtilizationScore has a similar relationship as CreditScore with the component features.

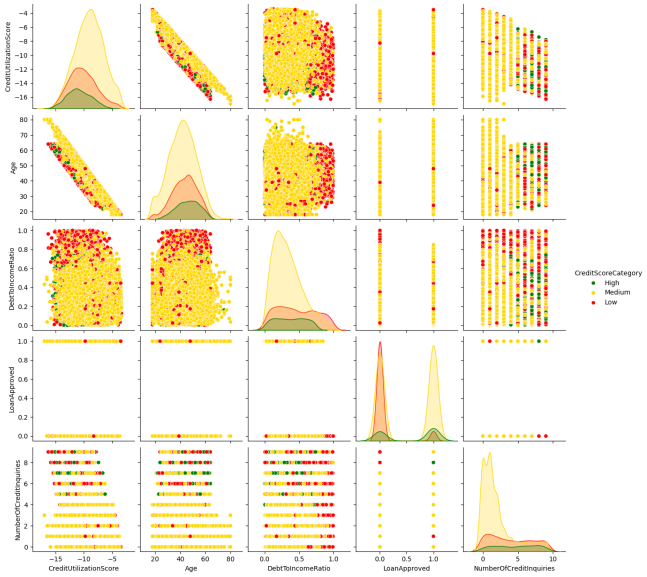

Fig. 19. Scatterplot matrix to explore the relationship of CreditUtilizationScore with its component factors

In Fig.20 we visualize a bar chart showing mean credit utilization score by loan approvals and we can see that the credit utilization score is more negative for people whose loan got denied than approved, similar to the way loan denied leads to lower credit scores as seen in Fig.13.

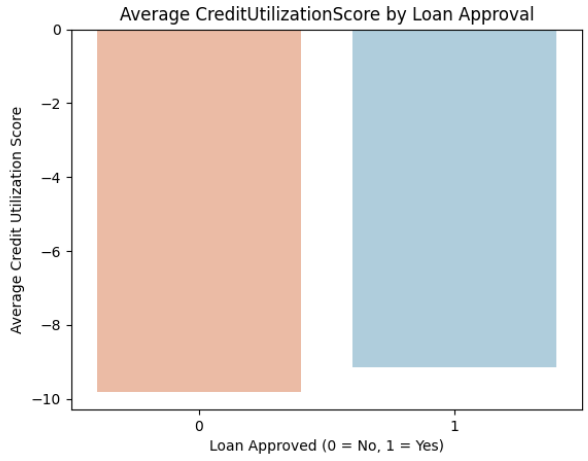

Fig. 20. Bar chart showing the mean value of CreditUtilizationScore across Loan Approval

In Fig.21, a combined violinplot showing the distribution of CreditUtilizationScore across NumberOfCreditInquiries, we

can see that CreditUtilizationScore becomes more negative as the NumberOfCreditInquiries increases similar to the way credit score is expected to decrease as the number of credit inquiries increases.

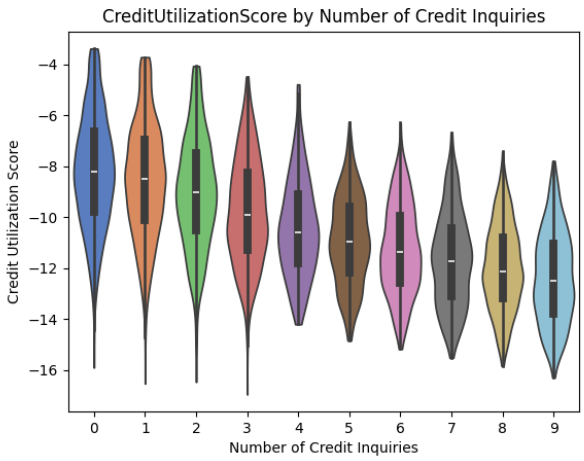

Fig. 21. Combined violinplot showing the distribution of CreditUtilization-Score across NumberOfCreditInquiries

A line plot as seen in Fig.22 shows the increase in negative magnitude of CreditUtilizationScore with an increase in Job Tenure, possibly indicating the availabilty of more credit lines and its subsequent use as one continues to remain in the same job over a long period of time.

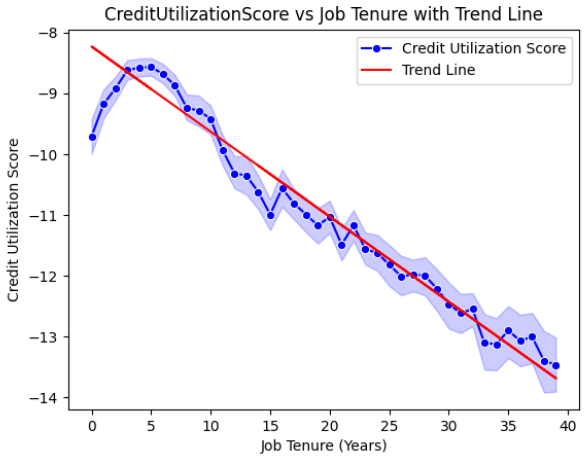

Fig. 22. Line plot showing the CreditUtilizationScore vs JobTenure

Now that we are done with visualizations, in iteration 2, we run a Gaussian Naive Bayes model again including the engineered feature, CreditUtilizationScore. The accuracy now comes out to be 74.75%, a 6.55% increase in accuracy compared to the previous model without the engineered feature. A side-by-side bar chart showing the F1 score of each credit score category across both models is shown in Fig.23, and it shows increase in accuracy over all credit score categories and therefore an overall increase in accuracy as well.

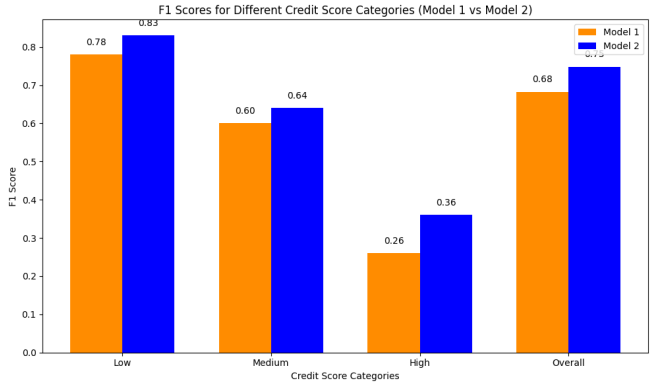

Fig. 23. Side-by-side Bar chart showing the accuracy across different credit score categories

The conclusion we can take from this is that our model with the engineered feature performs relatively well compared to the one without it. We can further improve the accuracy by considering some post-classification rules, one of which could be : if debt-to-income ratio is above 0.7 then the credit score category of that person cannot be high. So, we can forcefully tune the model outputs according to such rules. We can also consider other features instead of the one's used in the workflow to possibly get some new knowledge in the upcoming iterations of the workflow.

Workflow 3 - Analysing Behavioural Risk

Behavioural risk refers to the potential adverse impact on financial decisions or outcomes caused by human behaviour, biases, or psychological factors. These influences often lead individuals to act irrationally or against their best financial interests, which can increase financial risks, result in suboptimal decision-making, and amplify losses or missed opportunities. This workflow explores behavioural risk, compares it with traditional financial risk, and examines its implications for loan approval decisions. A conceptual sketch of the workflow is presented in Fig.24.

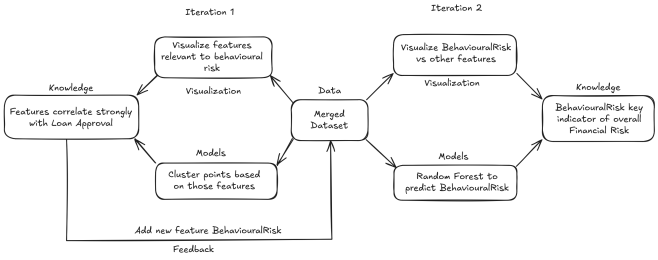

Fig. 24. Basic idea of the workflow

To that end, we begin by looking at some of the features in the dataset which are relevant to the financial behaviour and spending habits of people. These include features

such as *CreditCardUtilizationRate*, *DebtToIncomeRatio*, *SavingsAccountBalance*, *NumberOfOpenCreditLines* etc. Fig.25 and Fig.26 show treemaps of some of these features combined together, which give us an idea of the general demographics. It is immediately visible that a large amount of people have a high *CreditCardUtilizationRate* as well as a high *DebtToIncomeRatio*, which suggests that a significant section of the population is under financial strain. It is also inferred that a large majority of the borrowers are from the middle class, with a variety of loan amounts being borrowed, which is expected.

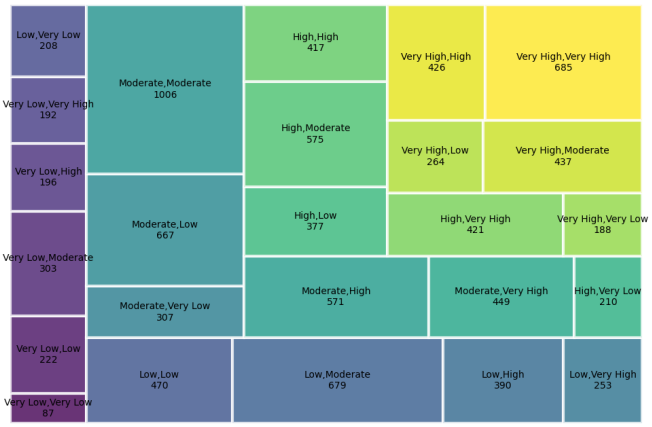

Fig. 25. Treemap of CreditCardUtilizationRate and DebtToIncomeRatio

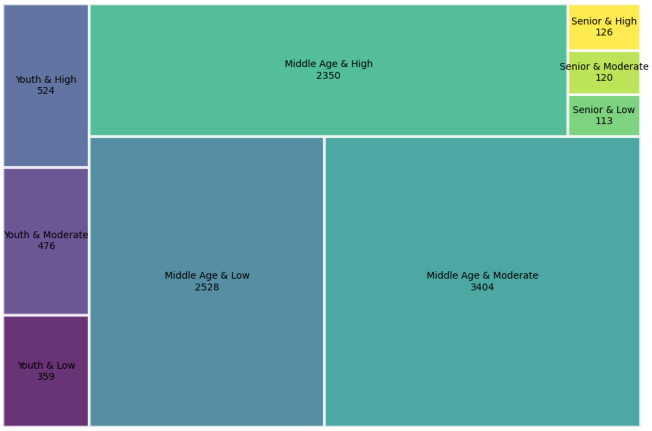

Fig. 26. Treemap of Age and LoanAmount

Next, we look at the distribution of various features and their combinations for unapproved loans. Fig.27 shows a pie chart with different categories of the ratio of *SavingsAccountBalance* and *AnnualIncome*. Most of the people who got their loan proposals rejected have around 10-20% of their annual income in their savings account. Fig.28 shows the number of open credit lines, which again is low (<5) for a majority of the population. A low credit activity is correlated with a low credit score, which can lead to loans not getting approved. Finally, Fig.29 shows the ratio of *TotalLiabilities* to *TotalAssets*. The pie chart is almost evenly split between low and high values, which seems to indicate that this ratio is not a huge factor in lending.

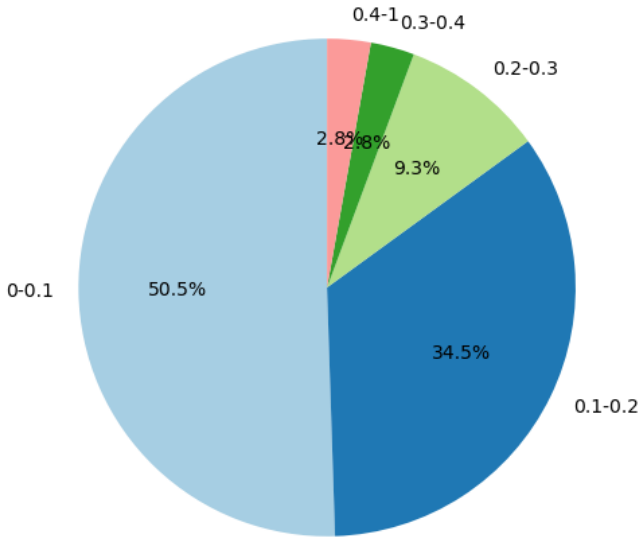

Fig. 27. Pie chart with Savings to Income Ratio for denied loans

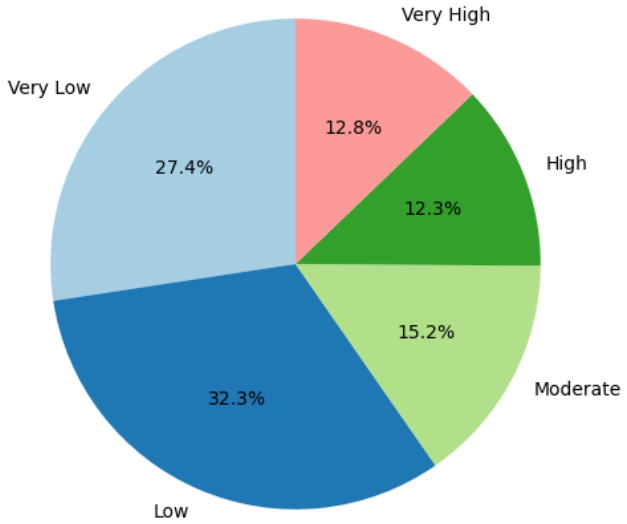

Fig. 28. Pie chart showing number of open credit lines for denied loans

Fig.30 shows a Sankey diagram which illustrates the flow of different features, and what values of what features lead to what approval status. A huge proportion of the unapproved loans are from the middle-age group, which itself is a huge

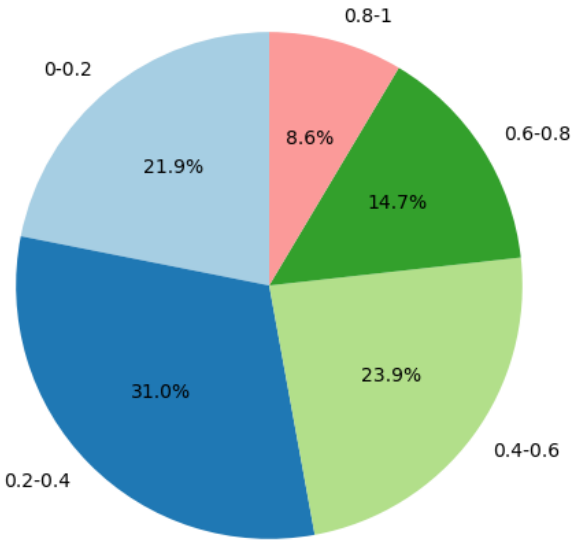

Fig. 29. Pie chart showing ratio of liabilities to assets for denied loans

proportion of the age groups applying for loans. Other than that, features like the loan purpose and education level are evenly distributed.

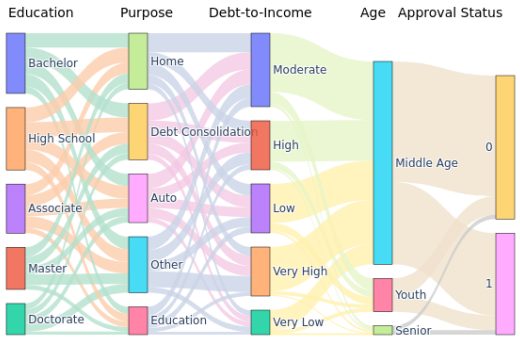

Fig. 30. Sankey diagram showing loan approval flow

Based on these visualizations, we can now create unsupervised clustering models which cluster data points based on a set of features. For instance, Fig.31 has 4 pie charts, which show the percentages of loans approved and not approved when the data is clustered in 4 clusters based on *CreditCardUtilizationRate* and *DebtToIncomeRatio*. We can see that more loans are approved for people with a low Debt-to-Income ratio and a low credit card utilization rate. If any one of these is high, the unapproved percentage shoots up. We can thus conclude that these are essential behavioural patterns to decide loan approval. A scatterplot of 50 randomly selected points is shown in Fig.32 to visualize the clusters.

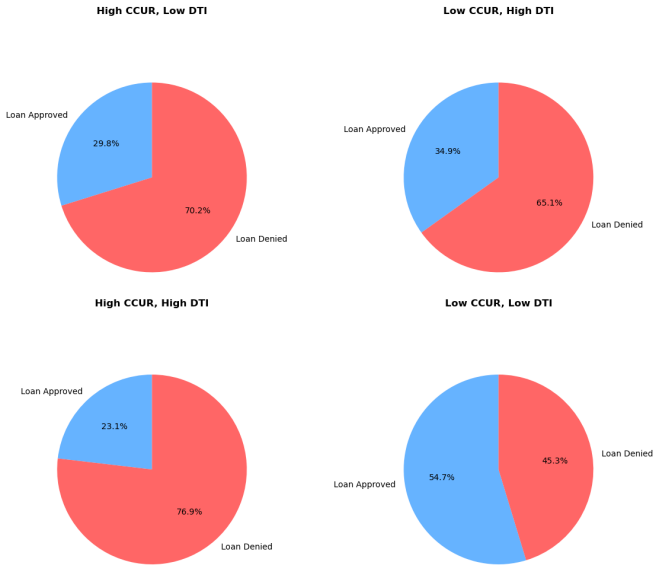

Fig. 31. Loan approval breakdown by debt to income ratio and credit card utilization rate

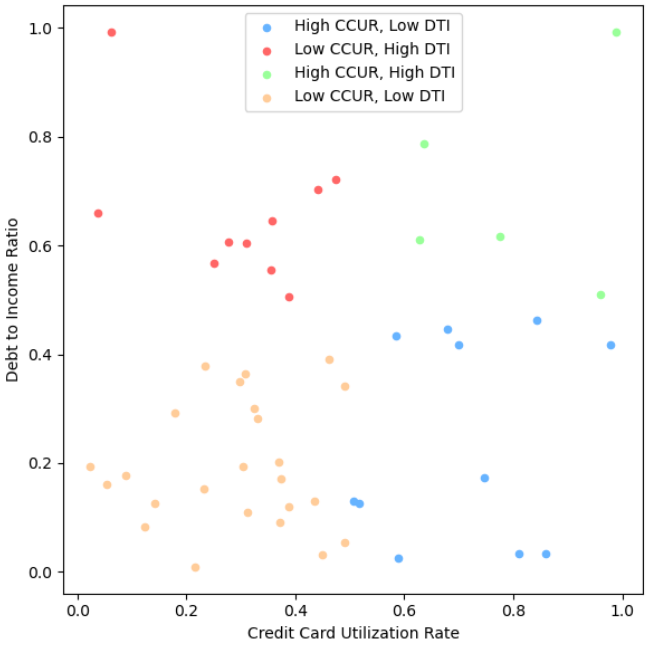

Fig. 32. Scatterplot showing clusters of loan approval breakdown by debt to income ratio and credit card utilization rate

Similarly, Fig.33 shows side-by-side bar plots to visualize 4 clusters, clustered based on *TotalAssets* and *TotalLiabilities*. In case of low assets, the amount of loans denied is higher irrespective of the amount of liabilities. When the amount of assets is higher than the amount of liabilities, more loans are approved. In case of equally high assets and liabilities, an

equal percentage of loans are approved. Note that overall, it looks like the majority of loans go unapproved.

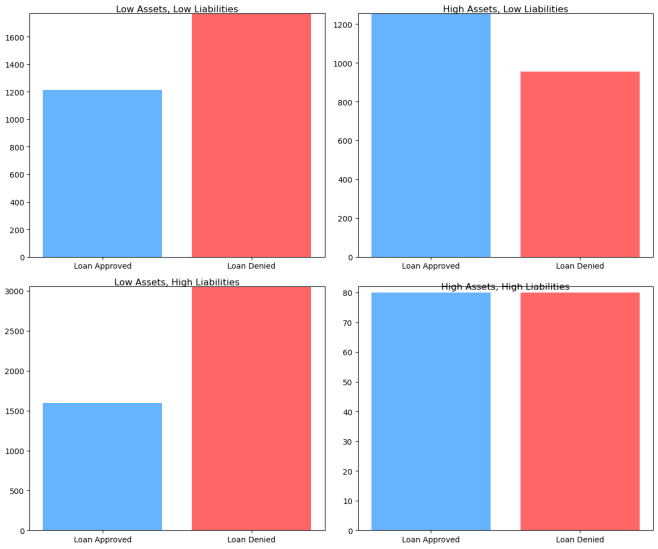

Fig. 33. Loan approval breakdown by assets and liabilities

Fig.34 shows a stacked bar plot which again visualizes 4 clusters. Here the data is clustered based on the savings-to-income ratio discussed earlier. It is clear that a low savings account balance directly contributes to a higher financial risk, as seen in the plot. On the other hand, higher savings do not directly indicate good financial health, and we also need to look at the annual income to better gauge the financial risk carried by the individual.

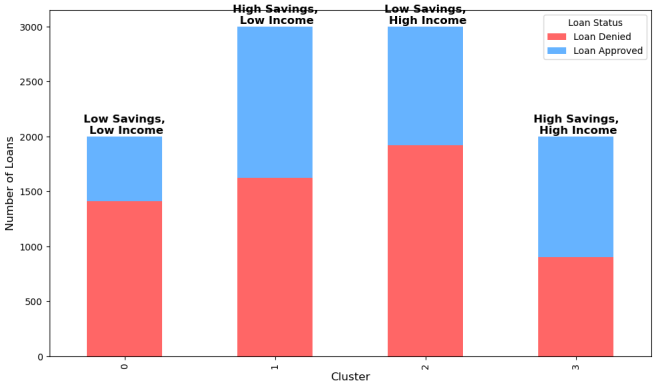

Fig. 34. Loan approval breakdown by annual income and savings account balance

The final clustering was carried out based on the *Age* and *LoanAmount* features, and we have a horizontal stacked bar plot with 6 bars as shown in Fig.35. We see a large proportion of middle-aged and young individuals applying for loans, as compared to the amount of senior citizens which is negligibly small, especially ones applying for bigger loans. Another thing to notice is that young people get bigger loans approved more easily, as they have lesser risks and a higher potential to repay the loans back in future without much hardships. On the other hand, smaller loans by young people often raise red flags about financial discipline, yield fewer returns to lenders, may be unsecured, may signify consumption rather than investment, and in general do not justify the operational costs and perceived risks brought by youth.

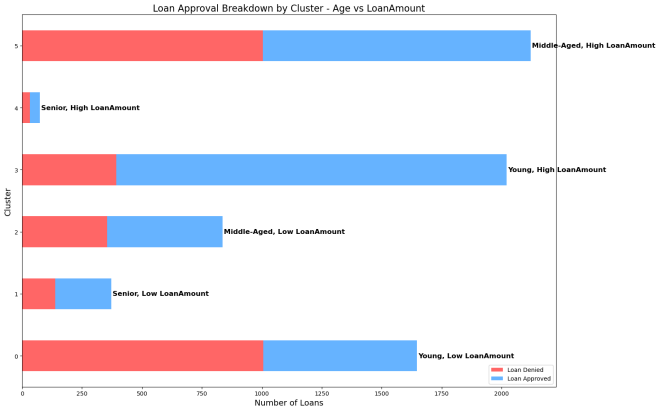

Fig. 35. Loan Approval breakdown by age and loan amount

Following are the general inferences we can get by looking at the above visualizations -

- There are a variety of factors affecting the loan approval status.
- Each of them may contribute positively or negatively to the perceived financial risk of a potential borrower.
- Patterns are observed when these factors are looked at in combination with other features.

We assume that these patterns are a reflection of real-life trends, and hence, by looking at these insights, we can suitably transform the data to augment our knowledge. We add a feedback loop i.e. data transformation as follows -

Add a feature called *BehaviouralRisk*, which is a weighted and normalized sum of the following:

- *CreditCardUtilizationRate*
- *DebtToIncomeRatio*
- *TotalLiabilities / TotalAssets*
- *Age*
- *LoanDuration*

Except the loan duration, an increase in any of these should typically lead to an increased behavioural risk, and we do observe that that is the case. Over 75% of the people with a behavioural risk score of over 0.7 have got their loans denied, while over 60% of the population with a score of less than 0.4 got their loans approved.

To check how the newly formed *BehaviouralRiskScore* compares with the traditional ways of assessing risk, like the credit score, payment history and number of open credit lines, we look at the following line plots -

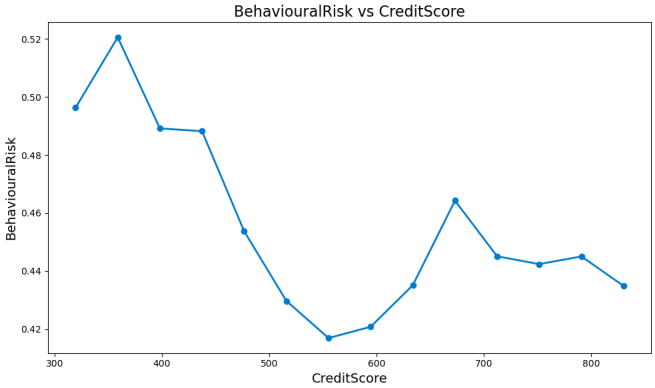

Fig. 36. Line chart of BehaviouralRisk vs mean CreditScore

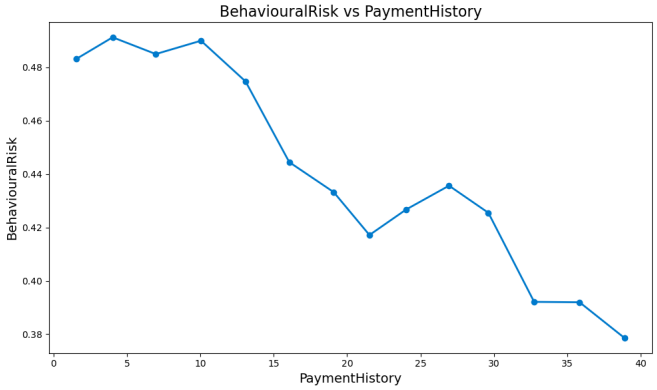

Fig. 38. Line chart of BehaviouralRisk vs PaymentHistory

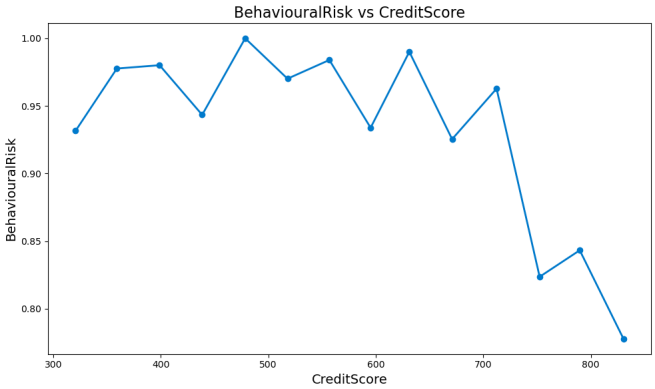

Fig. 37. Line chart of BehaviouralRisk vs max CreditScore

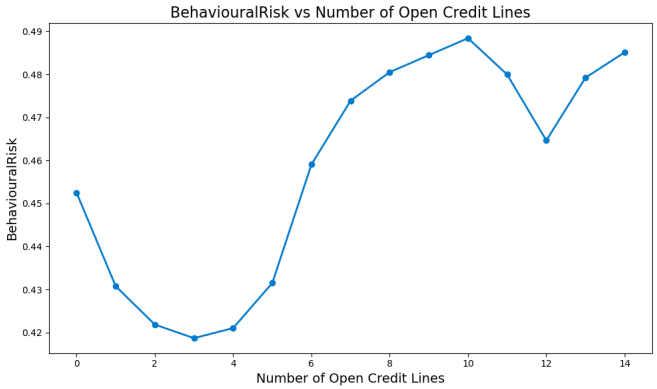

Fig. 39. Line chart of BehaviouralRisk vs NumberOfOpenCreditLines

In Fig.36, which shows the average behavioural risk score for a given credit score, we see that the behavioural risk drops with an increase in credit score, which is to be expected since a higher credit score directly corresponds to a lower risk of defaulting. The increase after a credit score of 600 is likely due to the age factor taken into account while computing behavioural risk, as it is generally unlikely for young people to have such a high credit score. Another thing to notice is that the mean behavioural risk lies somewhere around 0.4 and 0.5, which indicates that the average person is 'behaviourally' sufficiently credit-worthy. Fig.37 shows the same plot with the maximum behavioural risk score instead of the mean. There is no pattern here; the score keeps oscillating before finally at high credit scores.

Fig.38 and Fig.39 show how the behavioural risk score changes over payment history and the number of open credit lines. Again, the results are natural, because a longer payment history provides a more comprehensive record of the borrower's financial behaviour, while a large number of open credit lines indicates higher debt obligations. Even when we have not used these features for computing the behavioural risk, they still do affect it, because behavioural risk is directly correlated with the traditional model of measuring risk. The initial dip in the plot in Fig.39 can be explained by the fact that having a few open credit lines generally improves the credit score.

We then trained a simple random forest model to predict *BehaviouralRisk*, based on the columns not used in the formula. We got a validation set accuracy of over 80%, which indicates that the new feature we created is predictable using features in the dataset unrelated to its construction. In other words, a good estimate of a person's behavioural risk can be determined by looking at factors which may not directly contribute to, or come under the purview of, behavioural risk. Fig.40 shows the feature importances plot obtained by training the model. We can see that even though the relative importances are small, there are factors which contribute to the prediction, even though we would not think of them when we think of behavioural risk. Interest rate and annual income are good examples.

Fig.41 shows the effectiveness of using the behavioural risk as a threshold to approve/disapprove loans. A large number of people with a behavioural risk score of over 0.7 have got their loans rejected, while the majority of those who were below the threshold were 'safe'. Overall, we see that around 60% people did not get loans, even though we saw that the average behavioural risk score was around 0.45, which indicates that

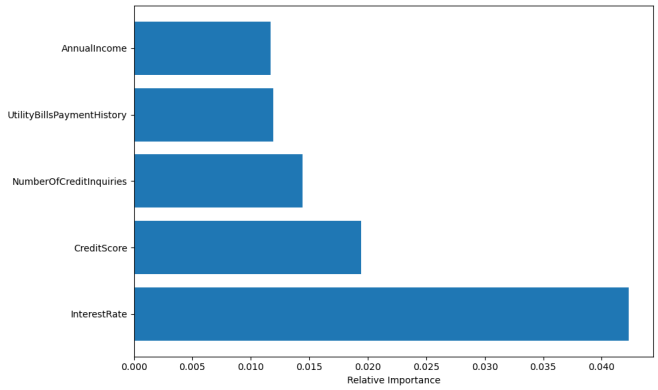

Fig. 40. Feature importances for BehaviouralRisk

a lot of people have risk scores above 0.7, but a lot of them also have extremely low risk scores, which pulls the average down. This can be easily confirmed - the median risk score of the population was 0.32, much less than the average - which means half the population had a very good score.

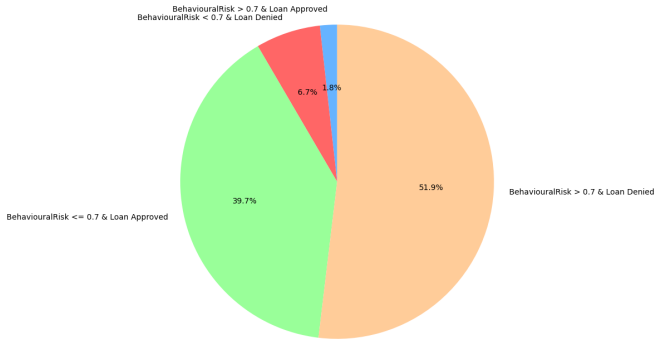

Fig. 41. Pie chart showing proportions of loan approvals based on BehaviouralRisk

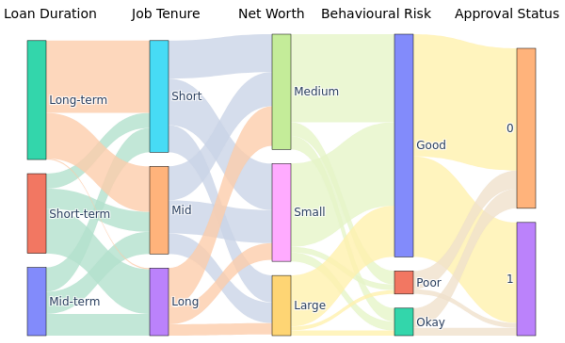

Fig. 42. Sankey Diagram of loan decision flow

Finally, we look at a Sankey diagram (Fig.42) to visualize the flow of loan approvals. We can see that a majority of the population takes long-term loans, are working in their current job for less than 15 years, have a small or medium net worth and have good financial behaviour. The loan approval status is almost evenly split between this class of individuals, along with some who are rich, and some who have taken short loans. Another thing to note is that most of the people who have a long job tenure have taken short or medium-term loans, which is not surprising since they must be old to have such a long tenure. Now, after looking at the clustering models, analysing and predicting the behavioural risk score and observing the associated visualization, we can draw the following inferences -

- There are a number of distinct patterns in the loan approval data i.e. certain segments of the population have higher approval/rejection rates than others, which was observed via the initial visualizations and unsupervised clustering.
- Introducing the *BehaviouralRisk* feature provides a lens to evaluate financial health and loan eligibility in a way different than what is done traditionally.
- Incorporating behavioral indicators into lending decisions enables a more nuanced and accurate risk assessment, while also being consistent with the given data.

Individual Contributions

- IMT2022017 Prateek Rath - Workflow 1 - Analysing model performance and behaviour on loan approval predictions
- IMT2022076 Mohit Naik - Workflow 3 - Analysing Behavioural Risk
- IMT2022103 Anurag Ramaswamy - Workflow 2 - Analysing Model performance and behaviour on credit score predictions

References

[1] D. A. Keim, F. Mansmann, and J. Thomas, "Visual analytics: how much visualization and how much analytics?" *Acm Sigkdd Explorations Newsletter*, vol. 11, no. 2, pp. 5–8, 2010.

[2] A. Powers, "Assignment 1 added as an appendix," GitHub, 2024, accessed: 2024-12-11.

[3] D. Mukherjee, "Financial risk data (large)," 2024. [Online]. Available: https://www.kaggle.com/datasets/deboleenamukherjee/financial-risk-data-large

[4] M. Faizan, "Financial risk for loan approval - ml ann," 2024. [Online]. Available: https://www.kaggle.com/code/muhammadfaizan65/financial-risk-for-loan-approval-ml-ann/input

[5] A. Hanbury, "Visual methods for analyzing probabilistic classification data," 2014. [Online]. Available: https://www.researchgate.net/publication/270789956_Visual_Methods_for_Analyzing_Probabilistic_Classification_Data

[6] B. Daniel, "Stacked bar charts efficacy," *Visual Informatics*, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468502X18300287