Introduce the problem

The bank is trying to speed up the loan approval decision process by observing features such as the customer's credit score, annual income, employment status, etc. The end goal is to have a time-efficient model that can accurately determine if a particular loan application will be approved.

1. I want to find which features influence the loan approval decision the most.
2. I want to find the best model with the highest accuracy score.
3. Is there any correlation between the various features?

Introduce the data

The dataset used for this project is:
https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset
I found it on Kaggle after browsing many datasets.
This dataset contains the following features:
Cibil_score- the applicant's credit score
Income_annum- the applicant's annual income
Loan_amount- the loan amount requested by the applicant
Loan_term- the loan term in years
Residential_assets_value- total residential assets (primary and rental properties) value of the applicant
Commercial_assets_value- total commercial (business) assets value
No_of_dependents- number of dependents (people financially dependent on applicant)
Education- whether the applicant is a graduate or not
Self_employed- are they self-employed?
And a couple more features.

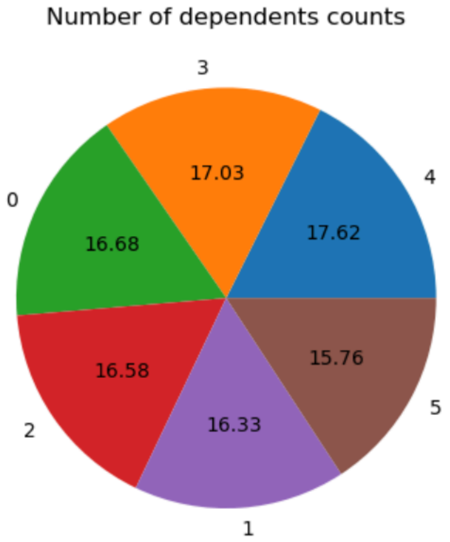
Preprocessing the data

1. Import libraries
2. Load the csv dataset file into a data frame
3. Check for null values in the dataset- there were no missing values
4. Understand the shape of the data- (4269,13)

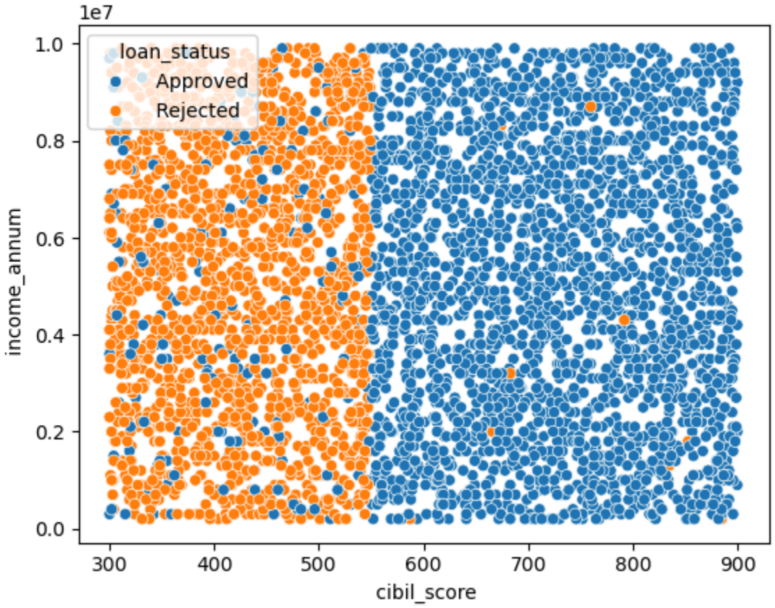
Data understanding/visualization

The dataset contains 9 features.
7 features are numeric, and 2 are non-numeric.
The shape of the data frame is 4269 rows by 13 columns.
Observations:

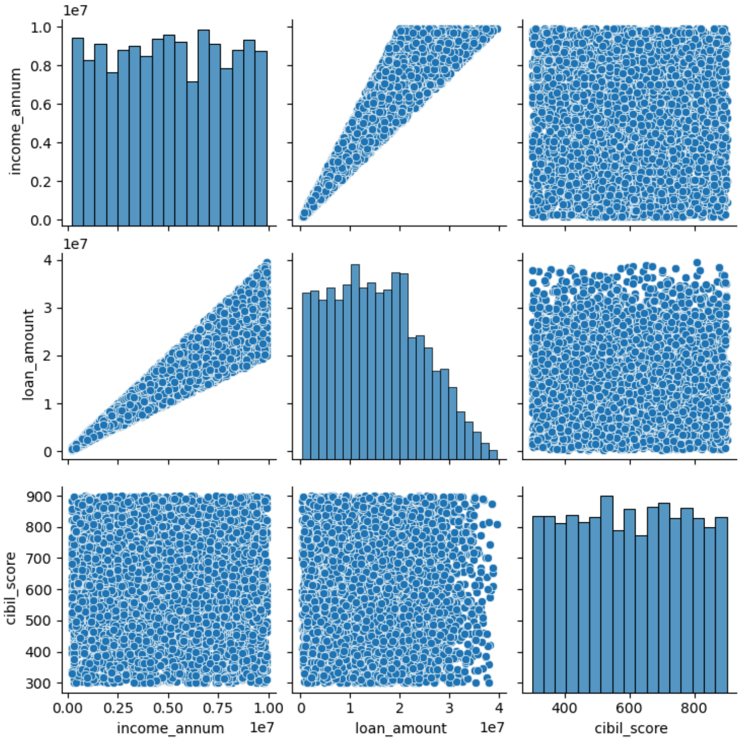
- You can see the highest and lowest annual income and the longest and shortest loan term requested.
- The same logic applies to columns: number of dependents, loan amount requested, residential assets value, commercial assets value, luxury assets value, and bank asset value.
- The data is evenly distributed among people with 0-5 dependents, graduate/not graduate, and loan term.



-
- The dataset contains more approvals than rejections (over 1000 more)
- Higher annual income barely influences loan approval; credit score does.
- The majority of people who requested for a loan and had a credit score of less than 550 were rejected.



-
- High loan amounts require a higher income, increasing the chance of approval. All three increase together.



-

Modeling
Three models: logistic regression, random forest, decision trees
Each model has strengths and weaknesses. I wanted to find the one that best fits for the classification problem I was trying to solve.
I referred to this article to look at them:
https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6
Logistic regression:
Pros- Simple
Cons- can easily find better results from other algorithms.
Results: Accuracy score was 73.88%
Decision trees:
Pros- selects relevant features automatically
Cons- sensitivity to changes in data
Result: accuracy score of 97.66%
Random forest classifier:
Pros- good with finding important features
Cons- requires features to have some predictive power
Result: accuracy score was 98.13%

Evaluation
Conclusion: Based on the model scoring evaluation, the random forest classifier model performed best for this problem. I used the model score to evaluate the accuracy.

Storytelling

Learnings:

The applicant's credit score is the highest influencer in a loan approval decision.

Features such as the number of dependents, education, and employment have very little impact on the decision.

For this project, I tried out different datasets and settled on this dataset because it has many records, making it suitable for training the models.

Impact

Positives-

One advantage of using these models is that they eliminate bias towards decisions.

It makes the approval process more time-efficient.

Negatives-

Could potentially eliminate manual jobs

References

https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset
https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6