# Assignment 1: Juan Escalada

## Introduction

In this report, we will be using loan application data to predict how much a given applicant will be approved for, given various attributes and the requested loan amount.

We will begin by exploring the data in the dataset, to get an idea of what a typical application looks like, as well as summarizing any patterns and correlations that may arise.

After that, we will create 3 models based on our analysis, and select the best model.

Our best model, along with our other models have a few variables which were highly significant: **Loan Amount Request (USD)**, **Co-Applicant** and **Credit Score** (specifically those in ranges 580-620, 620-650, 750-800, 800-850 and 850-900). These 3 factors almost single-handedly predict better than all the other features combined.

## Exploring the Data

We can start our data exploration by looking at our Predictor and Target variables.

**Predictor variables:** Gender, Age, Income (USD), Income Stability, Profession, Type of Employment, Location, Loan Amount Request (USD), Current Loan Expenses (USD), Expense Type 1, Expense Type 2, Dependents, Credit Score, No. of Defaults, Has Active Credit Card, Property ID, Property Age, Property Type, Property Location, Co-Applicant, Property Price

**Target variable:** Loan Sanction Amount (USD)

Based on these predictor variables, we can already make some assumptions in terms of which of these will be the most impactful: It is well-known that **Income, Credit Score and owning property are deciding factors in getting approved** for a loan, as well as accessing the best loan rates.

Let’s look at the correlations between some of these factors and the sanctioned amount (Fig. 1):

A graph of income and expenses

Description automatically generated with medium confidenceAs we predicted, the value of the property price, as well as the applicant’s credit score show correlations of over 0.5, making them very significant predictors of loan sanction amount. Unsurprisingly, the amount requested is also highly correlated to the amount sanctioned.

However, the applicant’s income and age, as well as whether they have a co-applicant, don’t seem to affect the approval amount.

Figure 1: Correlations between applicant attributes and approved loan amount

We may also want to look at the overall distribution of ages (Fig. 2), incomes (Fig. 3), property prices (Fig. 4) and loan sanction amounts (Fig. 5). There are a few patterns showing up:

1. There is a disproportionate number of 18- and 65-year-old applicants, which might indicate that the original loan application form only admitted values between 18 and 65. We would expect the ages to be more uniform. We also notice a slight increase in loan applications for ages 60+. **Median Age:** 40 ± 16 years
2. The incomes are fairly like a bell-curve, as expected. Some outliers exist, such as monthly incomes exceeding $30,000. Based on other characteristics of the application, we conclude that it is due to human error. These outliers make the variance high.

**Median Income (with outliers):** $2,225.29 ± $14,536.51

**Median Income (without outliers):** $2,223.6 ± $1,398.33

1. Property Prices are also fairly bell-curve shaped. We see that the typical value is less than $100,000.
2. The loan sanction amounts have a pronounced spike near the 0 mark. This represents all the loans that were rejected. We also notice that rather than a bell-curve, this is a bimodal distribution. There is a cluster of loans in the $15,000-$30,000 range, and another in the $40,000-$60,000 range.

A graph showing a green graph

Description automatically generated with medium confidenceA graph of a loan

Description automatically generated**Median Loan Sanction:** $34,695.19 ± $48,340.81

Figure 4: Property Price (USD)

Figure 5: Loan Sanction Amount (USD)

A graph with a bar graph

Description automatically generatedA graph of income and usd

Description automatically generated with medium confidence

Figure 2: Age distribution

Figure 3: Income distribution (USD per month)

We can now compare the most highly correlated application parameters to the sanctioned loan amount. By using scatterplots, we can see if there’s any explanation for the bimodal distribution in sanctioned loan amounts. We immediately see some interesting patterns regarding credit score (Fig. 6) and income (Fig. 7):

A graph showing a red line

Description automatically generatedA graph with a red line and a blue line

Description automatically generated

Figure 6: Credit Score vs. Loan Sanction Amount

Figure 7: Monthly Income vs. Loan Sanction Amount

The first thing we notice is that the minimum credit score for loan applications to be considered (save for some outliers) is 650. Meanwhile, those with scores over 850 always had their loans approved, regardless of loan amounts. Furthermore, there’s a direct positive relation with the sanctioned amount and the income, despite the initial correlation map indicating otherwise.

We can also look at how the value of the applicant’s property influences the sanctioned loan amount (Fig. 9). Those who don’t possess property are shown in the leftmost line on the y axis. These had a rejection rate of 13.4%. Among these, very few got approved for over $150,000. Those who possess property had an unexpectedly higher rejection rate of 28.0%, however their sanctioned loan amount increased linearly with the value of the property.

The applicant’s current loan expenses are also a large predictor of their sanctioned loan amount, which may seem paradoxical (Fig. 10). This may be due to applicants generally getting better credit scores as they borrow more money.

A graph showing a red line and a blue line

Description automatically generatedA graph showing a red line

Description automatically generated

Figure 10: Current Loan Expenses vs. Loan Sanction Amount

Figure 9: Property Price vs. Loan Sanction Amount

Based on our exploratory data analysis, we can conclude that property price, credit score, income and current loan expenses are the greatest predictors of sanctioned loan amounts. We will keep this in mind when developing our linear regression model.

## Developing Models

We developed various linear regression models to predict the sanctioned loan amounts of new applicants. One of them includes only the highly correlated factors we identified before. Two additional models have been implemented using recursive feature extraction (RFE). This will allow us to reduce bias in our models and find additional predictive factors.

### Model 1

This is a simple model which uses the four basic influential features we identified previously: Income, Loan Amount Request, Current Loan Expenses and Credit Score. All four of these are prime candidates for binning, however we will not perform it to keep things simple.

We implemented scaling to boost model performance. On top of that, invalid values have been replaced with 0.

### Model 2

This is another relatively simple model, which starts by generating dummy variables for every single categorical feature. These features are then filtered according to their F-statistics to see if they are relevant or not. We also performed clipping and imputing to fill in the missing values and increase performance.

### Model 3

This is a complete model which incorporates all the previously mentioned techniques and adds binning for the four main variables of interest: Credit Score, Income, Loan Amount Request, Current Loan Expenses.

The bins were selected conveniently according to the data we analyzed previously. Credit scores were grouped in groups of 50 (except for the 580-650 range which almost always gets rejected). Loan Amount Requests were grouped to aid recognition of the bimodal distribution identified before.

Furthermore, we have added an “Age 65” feature to encapsulate all those applicants with age over 65, and we used imputing to replace the ages of those aged 18. The reasoning behind this, is that those aged 18 and 65 make up a large percentage of the applications. Those aged 65 and over, get approved very often due to their steady income. The large number of 18-year-olds may be due to human error. The number is abnormally large, and all of them seem to have high incomes and possess property. Another explanation is that they are applying for a college loan with their parent as co-applicant. Regardless, their age should reflect the predicted age of their parent.

This time, we selected features manually to reduce the number of false positives based on f-statistics. Most of the features have an t-value greater than 2.

## Evaluating the Models

To evaluate the models, we use various metrics which are commonly used in machine learning, particularly for assessing Linear Regression models: Accuracy, Root Mean Squared Error, R2, R2adj, AIC, BIC.

Each model is cross-validated 5 times to achieve more accurate results for each of these metrics. Each metric has a mean and a standard deviation between all the generated cross-validated models.

The parameters for each model are shown in the diagram below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Mean (SD) | Parameters | Accuracy | RMSE | R2 | R2adj | AIC | BIC |
| Model 1 | 4 | 43.4%  (37.6%) | 35348.83 (8684.05) | 0.758 (0.001) | 0.757 (0.001) | 400316.58 (77.20) | 400347.53 (77.20) |
| Model 2 | 15 | 63.7% (1.3%) | 29122.52 (217.43) | 0.638 (0.003) | 0.637 (0.003) | 333199.89 (46.70) | 333313.35 (46.70) |
| Model 3 | 41 | 66.2% (2.0%) | 28091.08 (788.21) | 0.665 (0.005) | 0.664 (0.005) | 332143.72 (199.08) | 332438.72 (199.08) |

|  |  |
| --- | --- |
|  | Features (t-values > 15 highlighted) |
| Model 1 | Income (USD), **Loan Amount Request (USD)**, Current Loan Expenses (USD), Credit Score |
| Model 2 | Location\_Semi-Urban, Type of Employment\_Accountants, Profession\_Commercial associate, **Expense Type 1\_N,**  **Expense Type 1\_Y**, Income Stability\_Low, Type of Employment\_Laborers, Profession\_Working,  Type of Employment\_Managers, Location\_Urban, **Co-Applicant,** Credit Score, Current Loan Expenses (USD),  Property Price, **Loan Amount Request (USD)** |
| Model 3 | **Loan Amount Request (USD)**, Current Loan Expenses (USD), Dependents, No. of Defaults, **Co-Applicant**, Property Price,  Age 65, Income Stability (High), Income Stability (Low), Profession (Pensioner), Type of Employment (Drivers),  Type of Employment (Laborers), Type of Employment (Managers), Expense Type 1 (N), Expense Type 1 (Y),  Expense Type 2 (N), Expense Type 2 (Y), Property Location (Rural), Property Location (Semi-Urban),  Property Location (Urban), **Credit Score (580 to 620), Credit Score (620 to 650)**, Credit Score (650 to 700),  Credit Score (700 to 750), **Credit Score (750 to 800), Credit Score (800 to 850), Credit Score (850 to 900)**,  Loan Amount Request (0 to 15000), Loan Amount Request (15000 to 25000),  Loan Amount Request (25000 to 35000), Loan Amount Request (55000 to 80000),  Loan Amount Request (80000 to 125000), Loan Amount Request (125000 to 420000),  Current Loan Expenses (USD) (0 to 150), Current Loan Expenses (USD) (150 to 250),  Current Loan Expenses (USD) (250 to 350), Current Loan Expenses (USD) (350 to 450),  Current Loan Expenses (USD) (450 to 550), Current Loan Expenses (USD) (550 to 700),  Current Loan Expenses (USD) (700 to 1000), Current Loan Expenses (USD) (1000 to 1500) |

Each model has been cross-validated with k = 5. Increasing k results in lower RMSE with greater variance (potentially overfitting the model to the training and test data).

For our best model, we will also plot the Predicted vs. Actual values (Fig. 11) along with the error residuals representing the difference between predicted and actual values (Fig. 12).

A graph with blue dots

Description automatically generatedA graph with blue dots and a line

Description automatically generated

Figure 12: Residuals vs. Actual values for Model 3

Figure 11: Predicted vs. Actual values for Model 3

## Model Selection

Based on the model summary comparison, I selected Model 3. This one has the highest accuracy, as well as the lowest mean RMSE over multiple cross-validated runs. It occasionally reaches RMSE of 20000 or less, depending on the size of the cross-validation samples (which may be due to overfitting). The model’s R2 is 0.665, indicating a very strong correlation between predictions and actual values. However, as we can see in the previous charts (Figs. 11, 12), it sometimes fails to predict a rejection, which is ultimately the cause of the RMSE being relatively high. The model draws a line which is very close to the perfect values.

We could improve this model by using a classifier algorithm (such as logistic regression), which would allow us to first predict whether the applicant would get rejected in the first place. These could have a default loan value of 0.

Another improvement we could do to this model, is apply an extra function to classify the approval amount into one of 7 possible bins: 0 (rejected), 65%, 70%, 75%, 80%, 90%, 100%. According to our data, applicants don’t always get approved for the full amount they request, receiving instead a partial loan (Fig. 13).

A graph of a line graph

Description automatically generatedHere, we can see very clear lines at certain angles. The blue line represents the values where the Loan Requested is equal to the Loan Sanctioned. The other lines correspond to 90%, 80%, 75%, 70%, 65% and 0% (rejection) respectively.

By rounding our predictions to the nearest valid multiple, as well as predicting rejections beforehand, we could achieve much higher accuracy than we would through a simple linear regression model.

## Model Interpretation

The equation for our final model, is quite complicated as it has 41 independent variables:

*Loan Sanction Amount (USD) = -5676.3417 +* ***0.6114 \* Loan Amount Request (USD)*** *+ 18.6562 \* Current Loan Expenses (USD) - 757.1668 \* Dependents - 1396.4803 \* No. of Defaults +* ***29203.7244 \* Co-Applicant*** *- 0.0207 \* Property Price -3371.0064 \* Age 65* ***- 8313.1303 \* Income Stability\_High******- 6612.5372 \* Income Stability\_Low*** *+* ***12247.441 \* Profession\_Pensioner*** *- 1857.9061 \* Type of Employment\_Drivers -1253.6718 \* Type of Employment\_Laborers + 3350.7212 \* Type of Employment\_Managers - 3173.5803 \* Expense Type 1\_N - 2502.7613 \* Expense Type 1\_Y - 3314.8481 \* Expense Type 2\_N - 2361.4935 \* Expense Type 2\_Y - 6176.6418 \* Property Location\_Rural - 6036.8106 \* Property Location\_Semi-Urban - 5068.4333 \* Property Location\_Urban* ***- 29772.5351 \* Credit Score Bin\_(580, 620]******- 27603.4852 \* Credit Score Bin\_(620, 650]*** *+ 3215.2841 \* Credit Score Bin\_(650, 700] + 3631.6388 \* Credit Score Bin\_(700, 750]* ***+ 8514.5487 \* Credit Score Bin\_(750, 800] + 9647.3057 \* Credit Score Bin\_(800, 850] + 26690.9014 \* Credit Score Bin\_(850, 900]*** *+ 2531.6493 \* Loan Amount Request Bin\_(0, 15000] + 2667.7943 \* Loan Amount Request Bin\_(15000, 25000] + 2244.7277 \* Loan Amount Request Bin\_(25000, 35000] - 2441.9608 \* Loan Amount Request Bin\_(55000, 80000] - 4332.0744 \* Loan Amount Request Bin\_(80000, 125000] - 4607.6886 \* Loan Amount Request Bin\_(125000, 420000]* ***- 7948.5741 \* Current Loan Expenses (USD) Bin\_(0, 150] - 9696.8874 \* Current Loan Expenses (USD) Bin\_(150, 250] - 10783.2469 \* Current Loan Expenses (USD) Bin\_(250, 350] - 12323.2013 \* Current Loan Expenses (USD) Bin\_(350, 450] - 15257.2572 \* Current Loan Expenses (USD) Bin\_(450, 550] -17568.1113 \* Current Loan Expenses (USD) Bin\_(550, 700] - 22517.2451 \* Current Loan Expenses (USD) Bin\_(700, 1000] - 26794.5916 \* Current Loan Expenses (USD) Bin\_(1000, 1500]***

The requested loan amount is multiplied by around 0.6, which might indicate that people often get approved for loans smaller than what they ask for. Having a co-applicant increases the loan value by almost $30,000.

Pensioners are approved on average for $12,000 more than regular workers. Drivers and Laborers seem to be somewhat penalized, whereas Managers are more likely to get approved for larger loans.

Credit Score is by far the most influential feature: being in the worst credit score bins (580-620 or 62-650) reduces your predicted loan sanction by almost $30,000. On the other hand, those in the upper ranges (750-800, 800-850 and 850-900) seem to enjoy a bonus to their assigned amounts. Those in the best credit score bin receive on average an extra $30,000.

The sanctioned loan amount also depends on whether applicants ask for a little or a lot. Smaller loans of up to $35,000 are more likely to get approved, with a bonus factor of around $3,000. Larger loans over $55,000 are penalized by around $4,000.

Surprisingly, owning property in fact decreases the sanctioned loan amounts. The reason why is very unclear, as the property value should serve as collateral for a loan. Having any type of property reduced the loan sanctioned amounts by around $6,000. This penalty is further increased by property price. A property of price $200,000 would have a total penalty of $10,000.

Finally, another important factor deciding the loan amount, is the current loan expenses. In other words, the larger the existing debt obligations, the less money they are allowed to receive. For those with minimal debt, the amounts are penalized by around $9,000, but those with large debts are penalized by up to $27,000.

Because of the large number of penalties, a bulk of predictions are actually negative. This is because the model is a simple linear combination (polynomial) and cannot admit extra processing. Ideally we would constrain the predictions to values that make more sense.

Note: In order to use the above equation, you must replace the values for the categorical values with either 0 if False, or 1 if True.

## Summary

The most significant variables that affect the total loan given are **Loan Amount Request (USD)**, **Co-Applicant** and **Credit Score,** which shouldn’t be surprising. The other variables have a much smaller impact, however there are a few surprising cases that couldn’t have been found without our analysis, such as the negative impact of having a property, or working in certain professions.

We were also able to quantify the impact of each factor, as explained in our model interpretation.