Resit – Classification and Clustering

# Exploration

## Numerical data

First, to gain an overall understanding of our data, I will use df.describe().

Print(df.describe())

A picture containing text, file

Description automatically generated

From this, we can see that CreditStatus contains only 1s and 0s, with about 55% 1s to 45% 0s. We can also see that ExistingCreditsAtBank contains more than just 0s and 1s as it may have initially appeared.

All data has a minimum of 0 and a maximum of 1, with the exception of Age. Before continuing, Age must be normalised:

Text, letter

Description automatically generated

Df.describe() after normalisation:

A picture containing text, keyboard, computer, electronics

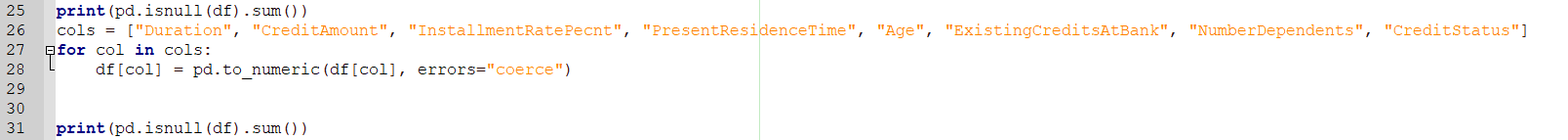
Description automatically generated

Age now has minimum = 0 and maximum = 1. However, this data will have to be normalised again. In particular, categorical data will need to be changed into numerical data and then normalized. Thus, I have written a more general function to normalise all of the columns in the data frame:

Graphical user interface, text

Description automatically generated

Next, we will start cleaning the data. First, we will run this code to find any null or non-numeric values within numeric columns:



Which prints:

Text

Description automatically generated

Table has no null values either side of running the above operations, thus there are no non-numeric values in numeric columns, so no data needs to be cleaned here.

Next, we shall both gain an understanding of the data and simultaneously identify outliers in our numerical data. Using a scatterplot and a linear regression function we can compare each datapoint to CreditStatus (our “y” or expected output) and see if there is a correlation. Using a boxplot and a histogram we can identify outliers to be removed later on (circled in red).

Chart, line chart

Description automatically generated

Duration vs. CreditStatus shows a fairly strong positive correlation, therefore Duration will likely be an important attribute for finding the creditworthiness of a customer.

Chart

Description automatically generated

Chart

Description automatically generated with medium confidence

CreditAmount has a strong correlation with CreditStatus and will therefore be an important attribute for finding the creditworthiness of a customer.

Chart

Description automatically generated

Chart

Description automatically generated

InstallmentRatePecnt has a weak positive correlation with CreditStatus and will therefore be less important than the previous two attributes. Still, it can potentially improve our algorithm, and will therefore be left in at this stage.

Chart

Description automatically generated

Chart

Description automatically generated

PresentResidenceTime has nearly no correlation will CreditStatus, suggesting there is no link between the two. Therefore, the PesentResidenceTime column will be dropped.

Chart

Description automatically generated

Graphical user interface, chart, line chart

Description automatically generated

Age has a strong negative correlation with CreditStatus. Negative correlations are as useful as positive correlations in determining potential causation. Thus, Age will likely be an important attribute for our model.

Chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

ExistingCreditsAtBank has a similar, slightly weaker negative correlation with CreditStatus. This is still important information for our logistic regression model and thus it will stay in our dataframe.

Chart, line chart

Description automatically generated

Chart

Description automatically generated

NumberDependents vs. CreditStatus shows nearly no correlation, and on top of this, shows NumberDependents contains more values than just 0 and 1 – whether these are invalid datapoints or not is hard to tell, but given that the attribute is documented as “numerical” not Boolean, I am going to assume they are valid. Given the lack of correlation, Number Dependents is likely going to be removed to improve the model.

Chart

Description automatically generated

## Non-numerical data

Chart, waterfall chart

Description automatically generated

A picture containing text

Description automatically generated

A11 and A12 have strong correlations with CreditStatus, A13 and A14 have strong negative correlations. Therefore, CheckingAccStat has strong correlations with CreditStatus and should be used in our logistical regression model.

Chart

Description automatically generated

A picture containing chart

Description automatically generated

A30, A31, A32, and A33 have positive correlations with CreditStatus, whilst A34 has a strong negative correlation with CreditStatus. Therefore, CreditHistory has strong correlations with CreditStatus and should be used in our logistical regression model.

Chart, bar chart, histogram

Description automatically generated

A picture containing text

Description automatically generated

Purpose has many categories, with varying correlations. Statistically insignificant categories may make the logistic regression model less accurate, but overall, the large categories with varying strong correlations outweigh this risk. Therefore, Purpose will be used in our logistical regression model.

Chart

Description automatically generated

Table

Description automatically generated with low confidence

A61 and A62 have strong positive correlations with CreditStatus, while the rest of the categories have negative correlations. Therefore, Savings has strong correlations with CreditStatus and should be used in our logistical regression model.

Chart, bar chart

Description automatically generated

A picture containing chart

Description automatically generated

Employment has a positive correlation with CreditStatus in all categories but one. As customers move from unemployed to employed to employed for many years, the correlation becomes weaker, even becoming negative in A74. Employment is correlated with CreditStatus in data and is intuitively linked to our output. Therefore, Employment will be used in our logistical regression model.

Chart

Description automatically generated

Text

Description automatically generated with medium confidence

All categories in sex and status have the same, fairly weak correlation with CreditStatus. Therefore, SexAndStatus will be dropped from our logistical regression model.

A picture containing waterfall chart

Description automatically generated

A picture containing text

Description automatically generated

From the histogram, we can see clearly that OtherDetorsGuarantors has 2 categories of close to no statistical significance. Furthermore, the frequency table comparing with CreditStatus shows only one of the categories shows a negative correlation. This would suggest that OtherDetorsGuarantors is not relevant to our problem and should be dropped.

Chart, bar chart

Description automatically generated

A picture containing text

Description automatically generated

Property has 4 categories with a relatively even spread of datapoints. The categories have varying correlations with CreditStatus, including some very strong correlations. Therefore, Property is a relevant column to our problem, and will be used in our logistical regression model.

Chart, bar chart

Description automatically generated

Text

Description automatically generated with low confidence

OtherInstalments has 3 categories, with one category being statistically significant. Furthermore, every category has a strong positive correlation with CreditStatus. Therefore, OtherInstalments has little to no effect on CreditStatus and is not relevant to our problem.

Chart

Description automatically generated

Text

Description automatically generated with medium confidence

Housing has three categories, all showing a positive correlation with CreditStatus. Although it may not be the most relevant datapoint, it is relevant enough to not be removed at this stage.

Chart, bar chart, histogram

Description automatically generated

Text

Description automatically generated with medium confidence

Job is intuitively relevant to Credit worthiness, but there is a very weak correlation between different Job categories and CreditStatus. As with housing, I cannot argue that Job should be removed as a relevant attribute at this stage.

A picture containing graphical user interface

Description automatically generated

A picture containing text

Description automatically generated

Intuitively, Telephone has little to no relevance to the Credit worthiness of a customer. To back this up, Telephone has almost no correlation with CreditStatus and should therefore not be used in our logistical regression model.

A picture containing text

Description automatically generated

A picture containing graphical user interface

Description automatically generated

There are only two categories in ForeignWorker. One of which is so small it could almost be disregarded as statistically insignificant. However, given that A202 has such a strong negative correlation with CreditStatus, it makes itself relevant enough to be considered.

## Categorical Encoding

Categorical data cannot be used in either logistic regression or for K-means clustering as these algorithms cannot aptly interpret Data such as “A121”, nor differentiate between this and “A122” in a meaningful way. Furthermore, “A121” cannot be used as a metric. Since it is not numerical, it makes no sense, for example, to find the distance between A121 and A92 in a K-means algorithm. Therefore, we must convert our categorical data into numerical data. We will do this in two ways: label encoding and one-hot encoding. Label encoding is converting data from categories into numbers. This is good in terms of reducing (or maintaining) the number of columns, but risks losing accuracy because the algorithm could weight arbitrary categories more than others. For example, if you were to use label encoding with the “Purpose” column, A40 or Car (new) would become 1, and A42 or furniture/equipment would become 3. But is it accurate to state that furniture is worth 3 times Car (new)? Not really. For this reason, we must only use label encoding when the data “categories” are somewhat continuous, such as with CheckingAcctStat. CheckingAcctStat has the following categories:

A11 : ... < 0 DM   
A12 : 0 <= ... < 200 DM   
A13 : ... >= 200 DM / salary assignments for at least 1 year   
A14 : no checking account

These are somewhat continuous categories: A11 is logically less than A12, which is less than A13. Thus, despite the small loss of accuracy this causes, we can use label encoding for these sorts of categories.

So, what to do with attributes like Purpose? One-hot encoding creates a new column for each option or category and uses a 1 to represent the presence of that category for each row, 0 otherwise. Below is a small example of what this looks like when done to Housing:

A picture containing text

Description automatically generated

Earlier in our exploration of data, I stated that SexAndStatus had no correlation and should therefore be dropped. After further consideration, I have decided to change SexAndStatus into Female (showing positive for female and negative for male datapoints) to determine if gender has an effect on our regression model. This has been done within our categorical\_encoding function, which also encodes all of our other categorical data:

Text

Description automatically generated with low confidence

Female shows a weak positive correlation with CreditStatus:Chart

Description automatically generated

The correlation is weak, but may still be useful for our logistic regression model. Thus, it will not be dropped.

# Logistic Regression Model

To gain a good understanding of how many features should be included in our logistic regression model, I ran through 2 features up to the number of columns in our dataframe:

Text

Description automatically generated with medium confidence

For the purposes of keeping the document at a reasonable length I will only show and discuss the key results, before deciding on the optimal number of features to be included.

## Evaluation

Chart, line chart

Description automatically generated

At 5 features, the AUC reaches 0.7.

Chart, line chart

Description automatically generated

At 7 features, the AUC reaches 0.75.

Chart, line chart

Description automatically generated

At 14 features, the AUC reaches 0.83. This was a considerable jump from 0.76 at 13 features. We may be reaching the optimal number of features here.

Chart, line chart, scatter chart

Description automatically generated

For 15-20 features, the AUC drops to 0.82, suggesting the optimal number of features was reached at 14.

Chart, line chart

Description automatically generated

The AUC remains the same up until 30 features, which includes all columns in our dataframe (not including the columns dropped manually). Thus, we have determined that our optimal number of features is 14. Here I have printed the f1-score along with other useful statistics about our model results:

A screenshot of a computer

Description automatically generated with medium confidence

We can see from this that we have an average accuracy of 74%

Our confusion matrix looks like this:

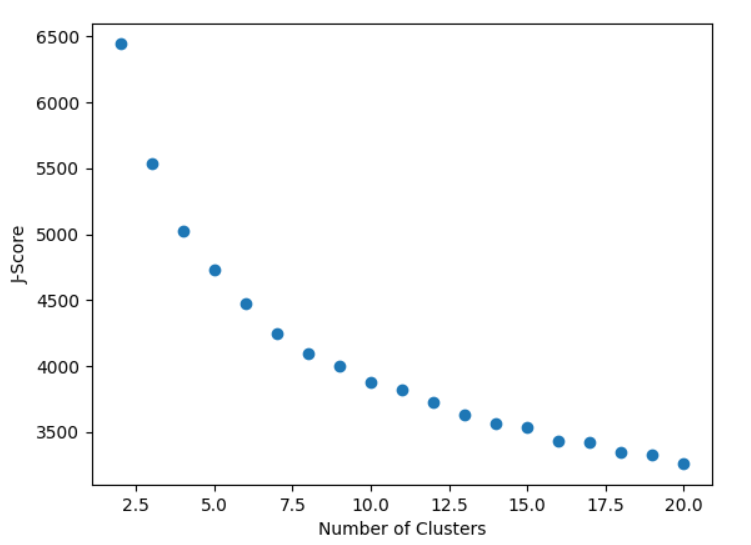
|  |  |  |
| --- | --- | --- |
| Actual/predicted | 0 | 1 |
| 0 | 16 | 126 |
| 1 | 5 | 48 |

Overall, I would say that our logistic regression model is decent, but 74% accuracy leaves a lot of room for improvement. From our confusion matrix, we can see that there are many false positives, which would be the main focus of any future improvements.

# Clusters

Using the below code, I have made a graph of the j-scores (cost) vs number of clusters in order to determine the optimal number of clusters utilizing the elbow method.Text

Description automatically generated



Elbow of the curve – 8 clusters

Chart, scatter chart

Description automatically generated

With 8 clusters:

Chart, scatter chart

Description automatically generated