# Changes

Saving datasets as “stage\_X\_data”

1. Added headers to the dataset (was 1000 rows with no headers previously)
2. Removed item ID numbers at the side

Normally in data science, will only lose data if necessary. (IBM?) Steps have taken in cases to write ML algorithms to try and update data rather than lose it

Sample – 1 row

Feature – 1 column

**Early on – checked good/bad split (700/300)**

**No female singles in dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column** | **Change from** | **Change to** | **Change No.** | **Why** |
| Added data | headers |  |  |  |
| Removed | IDs column |  |  |  |
| Job | yes | skilled | 2 | Assume that yes means skilled as skilled is overwhelmingly prevalent (62.8% of all applicants) and yes is a response to say “we’re skilled”  **If this were not skilled, it would be extremely important to classify since skilled overwhelmingly present** |
| Job | Removed ‘’s | - | 998 | Cleans up the data |
| Personal Status | ‘female div/dep/mar’ | female div/sep/mar | 310 | Cleaning data and removing typo |
| Personal Status | ‘male single’ | male single | 548 | Cleans up data |
| Personal Status | 'male mar/wid’ and 'male div/sep' | male mar/wid/div/sep | 92 mar/wid  50 div/sep  (Joined to 142) | Creates larger groups of data to check on, and if female data is already joined in group, makes sense for male too |
| Employment | Removed ‘’s | - | 938 | Cleans up the data |
| Saving Status | Removed ‘’s | - | 1000 | Cleans up the data |
| Saving Status | no known savings | unknown | 183 | More specific / clearer |
| Credit Requested | 111328000 | 13280 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 19280000 | 19280 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 13580000 | 13580 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 13860000 | 13860 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 5180000 | 5180 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 5850000 | 5850 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 7190000 | 7190 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 63610000 | 6361 | 1 | Looks like duplicated ‘1s’ and ‘0s’ (Changed down and not up since there are no other 5-figure loans within about £45,000 (next is £19,280). Meaning **crazy** data point |
| Purpose | ather | other | 1 | Wrongly typed |
| Purpose | busines | business | 3 | Wrongly typed |
| Purpose | Radio/Tv | radio/tv | 2 | Wrongly typed |
| Purpose | Eduction | Education | 1 | Wrongly typed |
| Purpose | busness | Business | 3 | Wrongly typed |
| Purpose | radio/Tv | radio/tv | 2 | Wrongly typed |
| Purpose | ‘domestic appliance’ | domestic appliance | 12 | Cleans up the data |
| Purpose | ‘used car’ | used car | 103 | Cleans up the data |
| Purpose | ‘new car’ | new car | 234 | Cleans up the data |
| Purpose | Merging domestic appliance and furniture/equipment | furniture/equipment | 12 (new field – size 193) | Merges the fields |
| Credit History | Removing ‘’s |  | 1000 | Cleans up the data |
| Credit History | Merging ‘no credits/all paid’ and ‘all paid’ | no credits/all paid | 49 (new field – size 89) | Merges the fields |
| Checking Status | Removing ‘’s |  | 1000 | Cleans up the data |
| Age | -29 | 29 | 1 | Updating error |
| Age | -34 | 34 | 1 | Updating error |
| Age | -35 | 35 | 1 | Updating error |
| Age | 0.44 | 44 | 1 | Updating error |
| Age | 0.24 | 24 | 1 | Updating error |
| Age | 0.35 | 35 | 1 | Updating error |
| Age | 6.0 | 60 | 1 | Updating error |
| Age | 222.0 | 22 | 2 | Updating error |
| Age | 1.0 | 25 | 1 | Updating - explain |
| Age | 333.0 | 33 | 1 | Updating error |

These calculations can also be based on other factors e.g. applicant, loan volume etc.

Didn’t mix education with retraining due to high volume of “good” in retraining and middle volume of “good” in education (28 good/22 bad) vs (8 good/1 bad) (56% good vs 89% good)

Haven’t merged new and used car either

Used car (86/17| 103) = 83%

New car (145/89| 234) = 62%

**Merging the following**

Domestic appliance (8/4 | 12) = 66%

Furniture/equipment (123/58 | 181) = 68%

radio/tv (218/62 | 280) = 78%

repairs (14/8 | 22) = 64%

retraining (8/1 | 9) = 89%

education (28/22 | 50) = 56%

Stages:

Stage 1 data: original

Stage 2 data: after first 2 changes (headers and job ‘’s)

Stage 3 data: after updates made to Credit Requested

Stage 4 data: after updates made to rest of table

Stage 5 data: ages updated

Talk about how I merged loads of fields and didn’t keep data backups, which started leading to sub-par training and a real difficulty of backing up data – meaning I had to redo

**Nominal:** Set lending criteria = 18-28, etc. because earliest age that lending can legally occur is 18

Try Increments of 5

Try 20-30 etc too

Mention that gradient boosting is super popular (AVAILABLE?)

Legally, when declining an application for credit, you are required to give a reason why the application was declined. Hence why NNs aren’t used. Meant I had to use a model which could show why the application was declined.

If a user defaults on a loan, it can cost the bank a lot of money. I placed more importance on trying to ensure enough applications were being declined. In some cases, there are more applications being accepted (True positives), but there was a large drop in declines. **Explain how bank makes money on loan**

Also talk about being competitive etc and opening themselves to risk (TN more important than TP

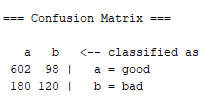
Use large sets for evaluation of rules to ensure max number of customers included

**ROC SUMMARY (**[**http://gim.unmc.edu/dxtests/roc3.htm**](http://gim.unmc.edu/dxtests/roc3.htm)**):**

* .90-1 = excellent (A)
* .80-.90 = good (B)
* .70-.80 = fair (C)
* .60-.70 = poor (D)
* .50-.60 = fail (F)

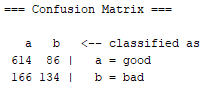
**Standard dataset (J48):**

* Accuracy: 72.2%
* Precision: 86.0%
* Recall: 77.0%
* F1 Score: 0.812
* ROC Area: 0.675



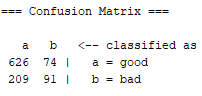
**Standard dataset (Random Forest):**

* Accuracy: 74.8%
* Precision: 78.7%
* Recall: 87.7%
* F1 Score: 0.830
* ROC Area: 0.764



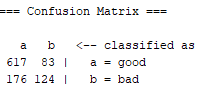
**18-23 age brackets (J48):**

* Accuracy: 71.7%
* Precision: 75.0%
* Recall: 89.4%
* F1 Score: 0.816
* ROC Area: 0.696



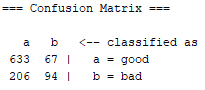
**18-23 age brackets (Random Forest):**

* Accuracy: 74.1%
* Precision: 77.8%
* Recall: 88.1%
* F1 Score: 0.827
* ROC Area: 0.758



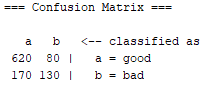
**18-28 age brackets (J48):**

* Accuracy: 72.7%
* Precision: 75.4%
* Recall: 90.4%
* F1 Score: 0.823
* ROC Area: 0.698



**18-28 age brackets (Random Forest):**

* Accuracy: 75.0%
* Precision: 78.5%
* Recall: 88.6%
* F1 Score: 0.832
* ROC Area: 0.752



**Random forest on standard dataset (possibly better than 18-28) and random forest on 18-28**

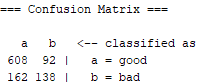
**Normalising and NN:**

Normalised lending by dividing by 20,000 – possibly safe to assume that lending over 20,000 shouldn’t be going through an automated system and should instead be checked by hand

Normalised age by dividing by 100

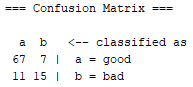
**Normalised data (Random Forest): - Best TN rate so far**

* Accuracy: 74.6%
* Precision: 79.0%
* Recall: 86.9%
* F1 Score: 0.827
* ROC Area: 0.762



**Normalised data (multi-layer perceptron) (2-layers, 1400 training time, 0.24 LR) (90:10 data split):**

* Accuracy: 82.0%
* Precision: 85.9%
* Recall: 90.5%
* F1 Score: 0.882
* ROC Area: 0.798

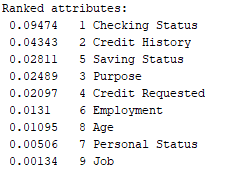


Since Male & Female mar/wid/div/sep showed very similar for NN weighting

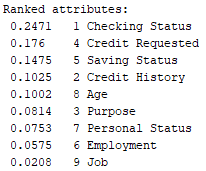
* Female div/sep/mar: -4.886645115129312
* Male mar/wid/div/sep: -4.

Attempting to join the two – Could be argued that creating ‘single’ and not male single allows single females to be factored into the model

**Original dataset (Correlation Attribute Evaluation):**

****

**Oversampled dataset (Correlation Attribute Evaluation):**



**Class imbalance:**

<https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/>

<https://content.pivotal.io/blog/how-to-deal-with-class-imbalance-and-machine-learning-on-big-data>

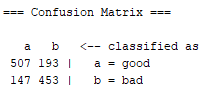
<https://www.einfochips.com/blog/addressing-challenges-associated-with-imbalanced-datasets-in-machine-learning/>

<https://www.researchgate.net/post/What_should_be_the_proportion_of_positive_and_negative_examples_to_make_a_training_set_result_in_an_unskewed_classifier>

<https://www.quora.com/I-have-an-imbalanced-dataset-with-two-classes-Would-it-be-considered-OK-if-I-oversample-the-minority-class-and-also-change-the-costs-of-misclassification-on-the-training-set-to-create-the-model/answer/Shehroz-Khan-2>

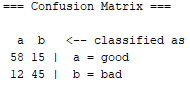
**Oversampled Dataset (J48) (Standard)**

* Accuracy: 73.8%
* Precision: 77.5%
* Recall: 72.4%
* F1 Score: 0.749
* ROC Area: 0.794



**Oversampled Dataset (J48) (Standard) (Percentage split 90:10)**

* Accuracy: 79.2%
* Precision: 82.9%
* Recall: 79.5%
* F1 Score: 0.811
* ROC Area: 0.804



Unemp/unskilled non res = unskilled (29)

Unskilled resident = unskilled (256)