Logistic Regression - Lasso 1

```
Full Dataset
[[22739   317]
 [  626   124]]

              precision    recall  f1-score   support

           0       0.97      0.99      0.98     23056
           1       0.28      0.17      0.21       750

    accuracy                           0.96     23806
   macro avg       0.63      0.58      0.59     23806
weighted avg       0.95      0.96      0.96     23806
```

```
Outliers Removed Dataset
[[17169    23]
 [  131     3]]

              precision    recall  f1-score   support

           0       0.99      1.00      1.00     17192
           1       0.12      0.02      0.04       134

    accuracy                           0.99     17326
   macro avg       0.55      0.51      0.52     17326
weighted avg       0.99      0.99      0.99     17326
```

The precision and recall for the full dataset are good for 0 and bad for 1. This is probably because of the imbalance.With the outliers removed the precision and recall are even worse, meaning that probably the outliers give some predictive power.

The macro average decreases with the removal of employers but weighted average increases. This might show that we have removed outliers from the minority class.

The F1 score is similarly good for 0 but bad for 1 and especially bad with outliers removed.

```
Resampled Dataset
[[9309  281]
 [ 791  377]]

              precision    recall  f1-score   support

           0       0.92      0.97      0.95      9590
           1       0.57      0.32      0.41      1168

    accuracy                           0.90     10758
   macro avg       0.75      0.65      0.68     10758
weighted avg       0.88      0.90      0.89     10758
```

The resampled dataset with outliers has better accuracy for 1 and a better weighted average overall. Similarly the dataset with outliers removed after resampling has better precision than the resampled dataset but worse precision for the minority class than the dataset with outliers.

```
Resampled Outliers Removed
[[7084   99]
 [ 623  216]]

              precision    recall  f1-score   support

           0       0.92      0.99      0.95      7183
           1       0.69      0.26      0.37       839

    accuracy                           0.91     8022
   macro avg       0.80      0.62      0.66     8022
weighted avg       0.89      0.91      0.89     8022
```

The F1 score - the measure of positive predictions which are correct - has increased a lot for the resample data.

Logistic Regression - Lasso 2

```
Full Dataset
[[22890   166]
 [  683    67]]
```

	precision	recall	f1-score	support
0	0.97	0.99	0.98	23056
1	0.29	0.09	0.14	750
accuracy			0.96	23806
macro avg	0.63	0.54	0.56	23806
weighted avg	0.95	0.96	0.96	23806

```
Outliers Removed Dataset
[[17187     5]
 [  134     0]]
```

	precision	recall	f1-score	support
0	0.99	1.00	1.00	17192
1	0.00	0.00	0.00	134
accuracy			0.99	17326
macro avg	0.50	0.50	0.50	17326
weighted avg	0.98	0.99	0.99	17326

```
Resampled Dataset
[[9235  355]
 [ 808  360]]
```

	precision	recall	f1-score	support
0	0.92	0.96	0.94	9590
1	0.50	0.31	0.38	1168
accuracy			0.89	10758
macro avg	0.71	0.64	0.66	10758
weighted avg	0.87	0.89	0.88	10758

```
Resampled Outliers Removed
[[7059  124]
 [ 651  188]]
```

	precision	recall	f1-score	support
0	0.92	0.98	0.95	7183
1	0.60	0.22	0.33	839
accuracy			0.90	8022
macro avg	0.76	0.60	0.64	8022
weighted avg	0.88	0.90	0.88	8022

With extra variables precision and recall are better here than above. Similarly, F1 score is better than above.

However the dataset with outliers removed has very bad precision and recall for the minority class. This is improved for the resample data set which has similar precision and recall to the logistic regression on only three variables.

Overall, three variables are likely too few, and therefore using more variables is probably a better idea.

Decision Tree

pred1

	precision	recall	f1-score	support
0	0.98	0.98	0.98	22915
1	0.42	0.36	0.39	891
accuracy			0.96	23806
macro avg	0.70	0.67	0.68	23806
weighted avg	0.95	0.96	0.96	23806

Already we see that scores are better than for the logistic regression models.

Similar to above, removing outliers from the unresampled data set reduces precision and recall in the 1 group.

pred2

	precision	recall	f1-score	support
0	0.99	0.99	0.99	17167
1	0.29	0.25	0.27	159
accuracy			0.99	17326
macro avg	0.64	0.62	0.63	17326
weighted avg	0.99	0.99	0.99	17326

In the resampled data, removing the outliers does not cause this effect. We might infer that a lot of the 1 class data points are outliers and are therefore being copied in the resampling process.

pred3

	precision	recall	f1-score	support
0	0.95	0.96	0.95	9500
1	0.64	0.59	0.62	1258
accuracy			0.91	10758
macro avg	0.79	0.77	0.78	10758
weighted avg	0.91	0.91	0.91	10758

The precision, recall and F1 score for the resampled dataset are quite good, meaning that we are predicting a lot of our minority class correctly, and we see this also in the macro and weighted averages.

pred4

	precision	recall	f1-score	support
0	0.96	0.97	0.97	7088
1	0.79	0.71	0.74	934
accuracy			0.94	8022
macro avg	0.87	0.84	0.86	8022
weighted avg	0.94	0.94	0.94	8022

Random Forest

forestpred1

	precision	recall	f1-score	support
0	1.00	0.98	0.99	23559
1	0.28	0.84	0.42	247
accuracy			0.98	23806
macro avg	0.64	0.91	0.70	23806
weighted avg	0.99	0.98	0.98	23806

forestpred2

	precision	recall	f1-score	support
0	1.00	0.99	1.00	17302
1	0.18	1.00	0.30	24
accuracy			0.99	17326
macro avg	0.59	1.00	0.65	17326
weighted avg	1.00	0.99	1.00	17326

forestpred3

	precision	recall	f1-score	support
0	0.98	0.96	0.97	9820
1	0.66	0.82	0.73	938
accuracy			0.95	10758
macro avg	0.82	0.89	0.85	10758
weighted avg	0.95	0.95	0.95	10758

forestpred4

	precision	recall	f1-score	support
0	1.00	0.98	0.99	7308
1	0.81	0.95	0.88	714
accuracy			0.98	8022
macro avg	0.90	0.97	0.93	8022
weighted avg	0.98	0.98	0.98	8022

The randomised benefit of a random forest is almost guaranteed to increase accuracy in predictions.

The 0 group is predicted almost perfectly, with the number of 1 group predicted accurately rising to 88% in the F1 score of the resampled dataset with outliers removed (forestpred4).

SVC

svc_preds1

	precision	recall	f1-score	support
0	1.00	0.97	0.98	23776
1	0.02	0.53	0.04	30
accuracy			0.97	23806
macro avg	0.51	0.75	0.51	23806
weighted avg	1.00	0.97	0.98	23806

The SVC has a hard time predicting the 1 group - probably because a hyperplane doesn't separate the data well.

Curiously, the precision for the 1 group is low in the unresampled data whereas the recall is high - meaning it is catching a lot of positives (1), but also getting a lot of false positives.

svc_preds2

	precision	recall	f1-score	support
0	1.00	0.99	1.00	17326
1	0.00	0.00	0.00	0
accuracy			0.99	17326
macro avg	0.50	0.50	0.50	17326
weighted avg	1.00	0.99	1.00	17326

svc_preds3

	precision	recall	f1-score	support
0	0.97	0.94	0.95	9933
1	0.46	0.66	0.54	825
accuracy			0.92	10758
macro avg	0.72	0.80	0.75	10758
weighted avg	0.93	0.92	0.92	10758

Precision increases for the 1 group in the resampled data (preds 3 and 4), but is still low compared to recall, meaning that we are still catching a lot of false positives - this means that the SVC is not a good model for this data.

svc_preds4

	precision	recall	f1-score	support
0	0.99	0.93	0.96	7653
1	0.33	0.74	0.45	369
accuracy			0.92	8022
macro avg	0.66	0.83	0.70	8022
weighted avg	0.96	0.92	0.93	8022

LDA

lda_preds1

	precision	recall	f1-score	support
0	0.97	0.98	0.98	22768
1	0.48	0.35	0.40	1038
accuracy			0.96	23806
macro avg	0.73	0.66	0.69	23806
weighted avg	0.95	0.96	0.95	23806

We see a similar situation for LDA that we did for SVC. Just as the data are not well-separated by a hyperplane, they are likely not well-separated by a linear boundary. However, here we have an opposite problem than with SVC, with our recall being lower than our precision for our unresampled data. This means we are predicting a lot of 0 group correctly, but not catching many 1 group data points.

lda_preds2

	precision	recall	f1-score	support
0	0.98	1.00	0.99	16907
1	0.46	0.15	0.22	419
accuracy			0.98	17326
macro avg	0.72	0.57	0.61	17326
weighted avg	0.97	0.98	0.97	17326

lda_preds3

	precision	recall	f1-score	support
0	0.98	0.95	0.96	9888
1	0.54	0.73	0.62	870
accuracy			0.93	10758
macro avg	0.76	0.84	0.79	10758
weighted avg	0.94	0.93	0.93	10758

For the resampled data, recall has increased - accurately predicting a lot more 1 group data points as positive. However, many accuracy measures are lower than the random forest models.

lda_preds4

	precision	recall	f1-score	support
0	1.00	0.95	0.98	7502
1	0.59	0.96	0.73	520
accuracy			0.95	8022
macro avg	0.79	0.96	0.85	8022
weighted avg	0.97	0.95	0.96	8022

QDA - all variables

qda_preds1

	precision	recall	f1-score	support
0	0.02	0.99	0.03	386
1	0.99	0.03	0.06	23420
accuracy			0.05	23806
macro avg	0.51	0.51	0.05	23806
weighted avg	0.98	0.05	0.06	23806

The precision for the 0 group is low - the QDA with all variables seems to be predicting all 0 group data points as 1-group.

This is seen also in the resampled dataset (preds 3 and 4). This is a strong sign that this model is not working well here.

qda_preds2

	precision	recall	f1-score	support
0	0.02	0.99	0.04	312
1	0.98	0.01	0.02	17014
accuracy			0.03	17326
macro avg	0.50	0.50	0.03	17326

qda_preds3

	precision	recall	f1-score	support
0	0.18	0.97	0.31	1812
1	0.95	0.12	0.22	8946
accuracy			0.27	10758
macro avg	0.57	0.55	0.26	10758
weighted avg	0.82	0.27	0.24	10758

qda_preds4

	precision	recall	f1-score	support
0	0.35	0.98	0.52	2551
1	0.95	0.15	0.25	5471
accuracy			0.41	8022
macro avg	0.65	0.57	0.38	8022
weighted avg	0.76	0.41	0.34	8022

QDA - Lasso 1 variables

qda_preds1

	precision	recall	f1-score	support
0	0.99	0.97	0.98	23378
1	0.18	0.31	0.23	428
accuracy			0.96	23806
macro avg	0.58	0.64	0.60	23806
weighted avg	0.97	0.96	0.97	23806

qda_preds2

	precision	recall	f1-score	support
0	0.99	0.99	0.99	17117
1	0.26	0.17	0.20	209
accuracy			0.98	17326
macro avg	0.63	0.58	0.60	17326
weighted avg	0.98	0.98	0.98	17326

qda_preds3

	precision	recall	f1-score	support
0	0.98	0.91	0.94	10393
1	0.18	0.58	0.28	365
accuracy			0.90	10758
macro avg	0.58	0.74	0.61	10758
weighted avg	0.96	0.90	0.92	10758

qda_preds4

	precision	recall	f1-score	support
0	0.99	0.92	0.95	7743
1	0.23	0.68	0.34	279
accuracy			0.91	8022
macro avg	0.61	0.80	0.65	8022
weighted avg	0.96	0.91	0.93	8022

Using the QDA with much fewer variables allows the model to predict the 0 group much more effectively.

However, precision for all data sets, even the resampled datasets with outliers removed (preds 3 and 4) is lower compared to the above models.

This means that 3 is probably too few variables to reliably predict on.

QDA - Lasso 2 variables

qda_preds5

	precision	recall	f1-score	support
0	0.96	0.98	0.97	22775
1	0.29	0.21	0.24	1031
accuracy			0.94	23806
macro avg	0.63	0.59	0.61	23806
weighted avg	0.94	0.94	0.94	23806

qda_preds6

	precision	recall	f1-score	support
0	0.94	1.00	0.97	16171
1	0.56	0.06	0.12	1155
accuracy			0.93	17326
macro avg	0.75	0.53	0.54	17326
weighted avg	0.91	0.93	0.91	17326

qda_preds7

	precision	recall	f1-score	support
0	0.96	0.92	0.94	9935
1	0.34	0.48	0.40	823
accuracy			0.89	10758
macro avg	0.65	0.70	0.67	10758
weighted avg	0.91	0.89	0.90	10758

qda_preds8

	precision	recall	f1-score	support
0	0.93	0.93	0.93	7198
1	0.42	0.43	0.43	824
accuracy			0.88	8022
macro avg	0.68	0.68	0.68	8022
weighted avg	0.88	0.88	0.88	8022

The QDA with the variables recommended by the second Lasso model gives better results than above, but not much better.

We found before that by manually selecting variables, we were able to get much more accuracy, although it was suspiciously good.

As the variables selected by Lasso 2 are only partially correlated with the binary response variable (because Lasso is a continuous method, and so we had to predict off of dep_inflow, not binary_response), this is likely the explanation for the lower performance of this model compared to the one we saw before with excellent accuracy.