# Avila Dataset

BDA 5002 – Marketing Analytics

06.05.2019

# The Dataset

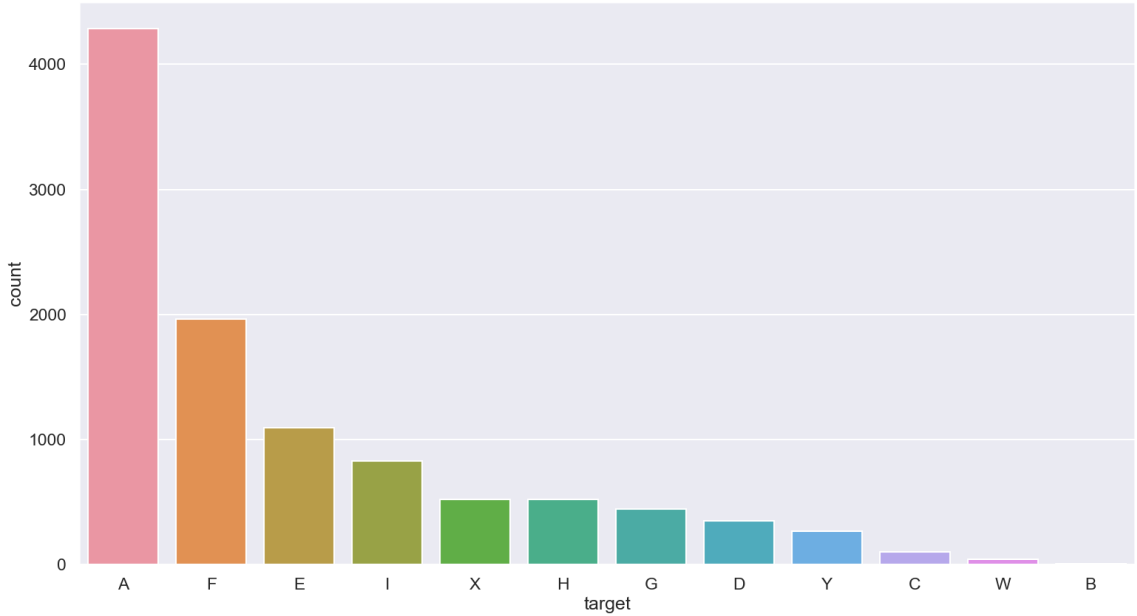
Sample data:

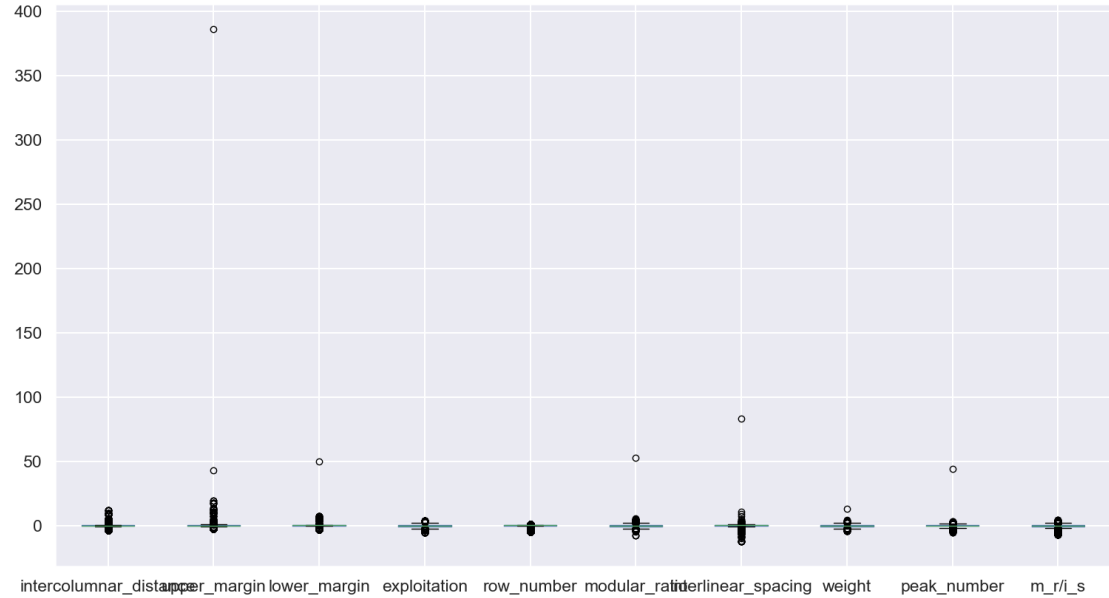
| | intercolumnar_distance | upper_margin | lower_margin | exploitation | \ |
|---|---|---|---|---|---|
| 0 | 0.266074 | -0.165620 | 0.320980 | 0.483299 | |
| 1 | 0.130292 | 0.870736 | -3.210528 | 0.062493 | |
| 2 | -0.116585 | 0.069915 | 0.068476 | -0.783147 | |
| 3 | 0.031541 | 0.297600 | -3.210528 | -0.583590 | |
| 4 | 0.229043 | 0.807926 | -0.052442 | 0.082634 | |

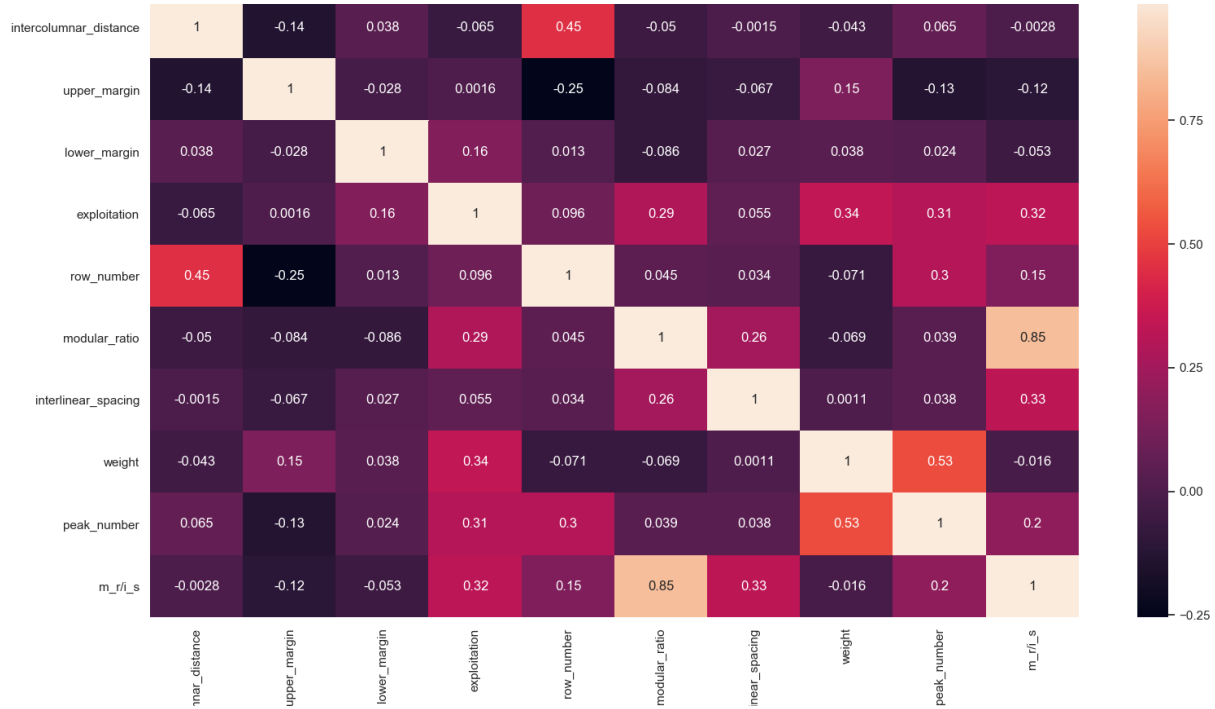
| | row_number | modular_ratio | interlinear_spacing | weight | peak_number | \ |
|---|---|---|---|---|---|---|
| 0 | 0.172340 | 0.273364 | 0.371178 | 0.929823 | 0.251173 | |
| 1 | 0.261718 | 1.436060 | 1.465940 | 0.636203 | 0.282354 | |
| 2 | 0.261718 | 0.439463 | -0.081827 | -0.888236 | -0.123005 | |
| 3 | -0.721442 | -0.307984 | 0.710932 | 1.051693 | 0.594169 | |
| 4 | 0.261718 | 0.148790 | 0.635431 | 0.051062 | 0.032902 | |

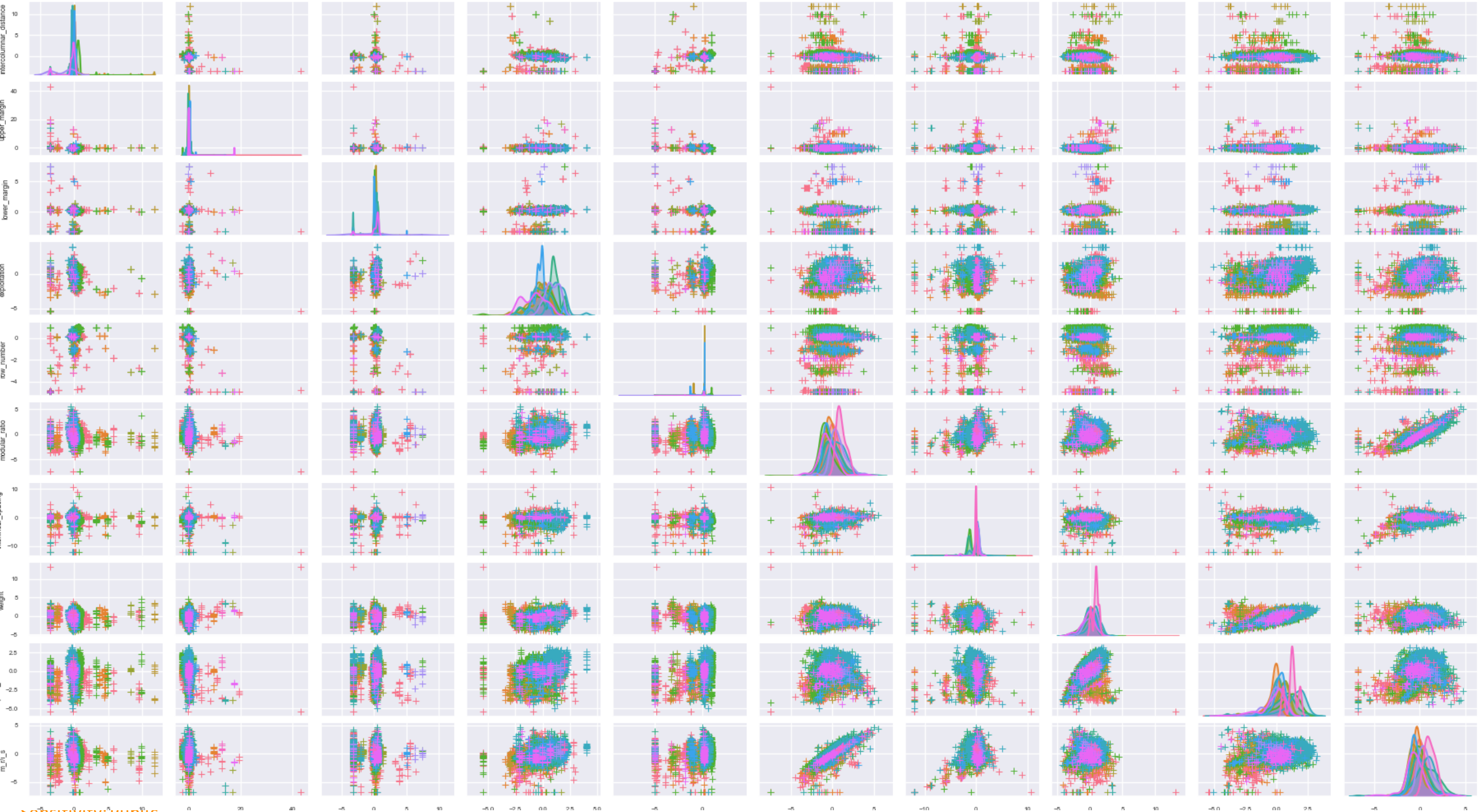
| | m_r/i_s | target |
|---|---|---|
| 0 | 0.159345 | A |
| 1 | 0.515587 | A |
| 2 | 0.582939 | A |
| 3 | -0.533994 | A |
| 4 | -0.086652 | F |

Number of rows in the dataset: 10430





# Feature Engineering

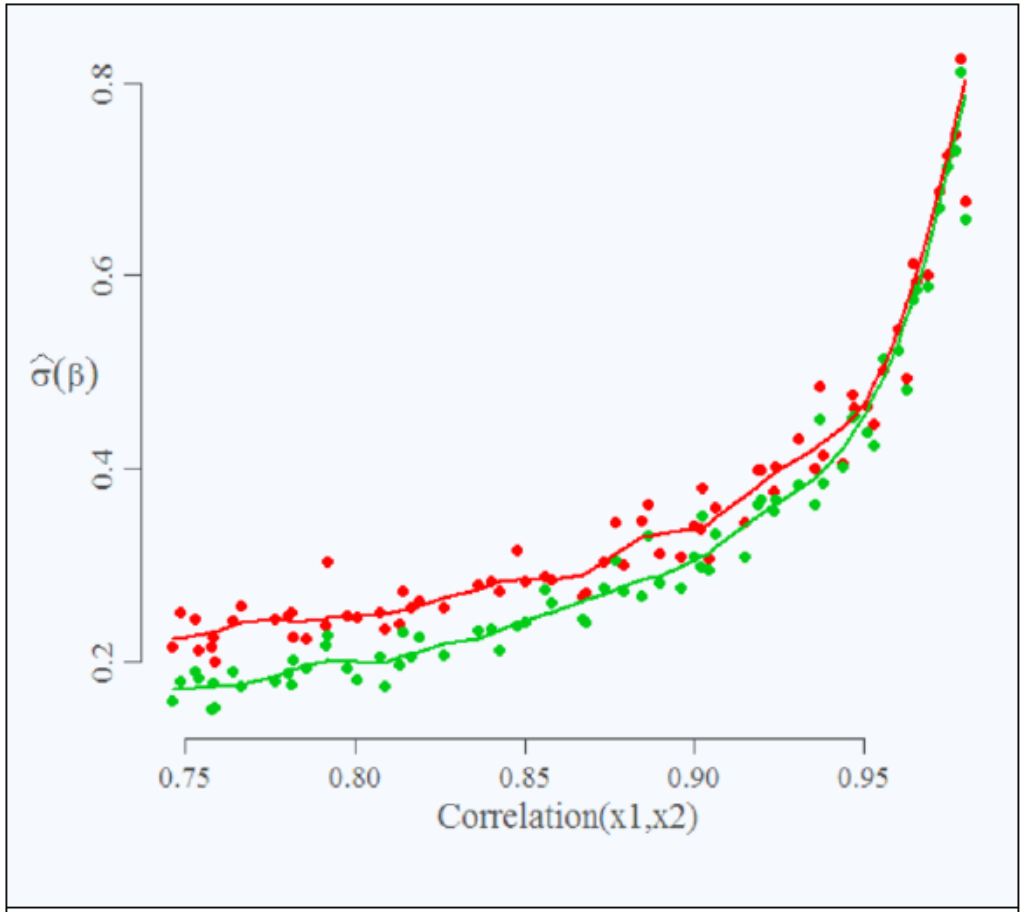
- One outlier data point is removed
- Data is split into training and test sets
- One feature is removed to prevent **multicollinearity** using VIF
- SMOTE (Synthetic Minority Over-sampling Technique) is used because of class imbalance problem

```
m_r/i_s is dropped with VIF=  4.22

Remaining features with VIF < 4:
intercolumnar_distance    1.29
upper_margin              1.15
lower_margin              1.05
exploitation              1.37
row_number                1.55
modular_ratio             1.26
interlinear_spacing       1.09
weight                    1.74
peak_number               1.74
```

$$VIF(\hat{\beta}_i) = \frac{1}{1 - R_i^2}$$

where $R_i^2$ is the squared multiple correlation coefficient between $x_i$ and the other explanatory variables.




You can see that even if correlation is around 0.8, inference is limited but perhaps manageable. As we move up in correlation, the standard deviation of the estimate is getting ridiculously high, in a sense that the estimate is 1, but it can be quite easily estimated as 0 or 2. This is shown by the increased standard deviation of the estimate, especially the sharp increase when correlation is above 0.95.

*Source: https://eranraviv.com/understanding-multicollinearity/*

# Algorithms' Performance on Training Set with 3-fold Cross Validation

## Logistic Regression with hyperparameter optimization

**Accuracy: 0.574 (+/- 0.017 with 95% CI)**

Accuracy: 0.574 (+/- 0.017)

| Predicted | A | B | C | D | E | F | G | H | I | W | X |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Actual | | | | | | | | | | | |
| A | 817 | 0 | 221 | 298 | 170 | 537 | 389 | 122 | 56 | 254 | 43 |
| B | 0 | 2999 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 35 | 0 | 1128 | 92 | 313 | 2 | 131 | 1083 | 28 | 150 | 7 |
| D | 346 | 0 | 223 | 958 | 396 | 110 | 146 | 405 | 120 | 141 | 7 |
| E | 115 | 0 | 271 | 161 | 830 | 164 | 308 | 378 | 23 | 208 | 346 |
| F | 506 | 0 | 217 | 208 | 103 | 855 | 500 | 403 | 44 | 110 | 0 |
| G | 214 | 0 | 120 | 6 | 22 | 176 | 1403 | 814 | 0 | 140 | 59 |
| H | 96 | 0 | 156 | 12 | 278 | 147 | 601 | 1533 | 69 | 5 | 48 |
| I | 23 | 0 | 72 | 0 | 5 | 38 | 9 | 6 | 2624 | 49 | 65 |
| W | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2999 | 0 |
| X | 10 | 0 | 7 | 37 | 12 | 5 | 13 | 121 | 64 | 99 | 2317 |
| Y | 42 | 0 | 21 | 44 | 50 | 3 | 39 | 79 | 115 | 228 | 185 |
| All | 2204 | 2999 | 2436 | 1816 | 2179 | 2037 | 3539 | 4944 | 3143 | 4383 | 3077 |

| Predicted | Y | All |
|---|---|---|
| Actual | | |
| A | 92 | 2999 |
| B | 0 | 2999 |
| C | 30 | 2999 |
| D | 147 | 2999 |
| E | 195 | 2999 |
| F | 53 | 2999 |
| G | 45 | 2999 |
| H | 54 | 2999 |
| I | 108 | 2999 |
| W | 0 | 2999 |
| X | 314 | 2999 |
| Y | 2193 | 2999 |
| All | 3231 | 35988 |

## AdaBoost Classifier (without SMOTE)

**Accuracy: 0.487 (+/- 0.024 with 95% CI)**

Accuracy: 0.487 (+/- 0.024)

| Predicted | A | B | D | I | W | X | Y | All |
|---|---|---|---|---|---|---|---|---|
| Actual | | | | | | | | |
| A | 2938 | 4 | 2 | 5 | 28 | 20 | 2 | 2999 |
| B | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 4 |
| C | 69 | 2 | 0 | 0 | 0 | 1 | 0 | 72 |
| D | 244 | 2 | 0 | 0 | 0 | 0 | 0 | 246 |
| E | 761 | 1 | 0 | 1 | 2 | 2 | 0 | 767 |
| F | 1368 | 0 | 0 | 0 | 2 | 3 | 0 | 1373 |
| G | 312 | 0 | 0 | 0 | 0 | 0 | 0 | 312 |
| H | 363 | 0 | 0 | 0 | 0 | 0 | 0 | 363 |
| I | 59 | 0 | 0 | 523 | 0 | 0 | 0 | 582 |
| W | 30 | 0 | 0 | 0 | 1 | 0 | 0 | 31 |
| X | 324 | 0 | 0 | 0 | 0 | 3 | 38 | 365 |
| Y | 96 | 0 | 0 | 0 | 0 | 3 | 87 | 186 |
| All | 6564 | 9 | 6 | 529 | 33 | 32 | 127 | 7300 |

## with SMOTE

**Accuracy: 0.279 (+/- 0.044 with 95% CI)**

Accuracy: 0.279 (+/- 0.044)

| Predicted | A | B | C | D | E | F | G | H | I | W | X | Y | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Actual | | | | | | | | | | | | | |
| A | 158 | 1 | 1 | 70 | 0 | 3 | 2271 | 0 | 26 | 328 | 121 | 20 | 2999 |
| B | 2100 | 0 | 300 | 599 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2999 |
| C | 61 | 7 | 19 | 85 | 9 | 0 | 2364 | 0 | 5 | 350 | 97 | 2 | 2999 |
| D | 7 | 2 | 3 | 61 | 0 | 0 | 2504 | 0 | 5 | 335 | 73 | 9 | 2999 |
| E | 28 | 0 | 0 | 41 | 0 | 0 | 2511 | 0 | 1 | 318 | 92 | 8 | 2999 |
| F | 66 | 0 | 0 | 20 | 0 | 1 | 2529 | 1 | 7 | 302 | 62 | 11 | 2999 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 2580 | 0 | 0 | 354 | 61 | 4 | 2999 |
| H | 54 | 0 | 0 | 10 | 0 | 3 | 2523 | 5 | 32 | 280 | 83 | 9 | 2999 |
| I | 104 | 0 | 0 | 39 | 0 | 0 | 0 | 0 | 2702 | 8 | 23 | 123 | 2999 |
| W | 110 | 0 | 0 | 858 | 0 | 0 | 250 | 0 | 0 | 1598 | 183 | 0 | 2999 |
| X | 9 | 0 | 0 | 26 | 0 | 0 | 1466 | 0 | 8 | 496 | 539 | 455 | 2999 |
| Y | 18 | 0 | 0 | 54 | 0 | 0 | 23 | 0 | 66 | 15 | 438 | 2385 | 2999 |
| All | 2715 | 10 | 323 | 1863 | 9 | 7 | 19021 | 6 | 2852 | 4384 | 1772 | 3026 | 35988 |

## Random Forests Classifier

**Accuracy: 0.997 (+/- 0.004 with 95% CI)**

Accuracy: 0.997 (+/- 0.004)

| Predicted | A | B | C | D | E | F | G | H | I | W | X |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Actual | | | | | | | | | | | |
| A | 2985 | 0 | 0 | 2 | 2 | 3 | 2 | 5 | 0 | 0 | 0 |
| B | 0 | 2999 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 2999 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 2999 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E | 7 | 0 | 0 | 4 | 2985 | 1 | 0 | 1 | 0 | 0 | 1 |
| F | 27 | 0 | 0 | 0 | 3 | 2937 | 21 | 11 | 0 | 0 | 0 |
| G | 1 | 0 | 0 | 0 | 1 | 1 | 2996 | 0 | 0 | 0 | 0 |
| H | 0 | 0 | 1 | 0 | 3 | 1 | 3 | 2991 | 0 | 0 | 0 |
| I | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2998 | 0 | 0 |
| W | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2999 | 0 |
| X | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 2988 |
| Y | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 5 |
| All | 3020 | 2999 | 3000 | 3007 | 3001 | 2943 | 3022 | 3008 | 2998 | 2999 | 2994 |

| Predicted | Y | All |
|---|---|---|
| Actual | | |
| A | 0 | 2999 |
| B | 0 | 2999 |
| C | 0 | 2999 |
| D | 0 | 2999 |
| E | 0 | 2999 |
| F | 0 | 2999 |
| G | 0 | 2999 |
| H | 0 | 2999 |
| I | 0 | 2999 |
| W | 0 | 2999 |
| X | 5 | 2999 |
| Y | 2992 | 2999 |
| All | 2997 | 35988 |

# Random Forests Classifier on Test Set with 3-fold CV

- **Accuracy: 0.957 (+/- 0.024 with 95% CI)**

Accuracy: 0.957 (+/- 0.024)

| Predicted | A | C | D | E | F | G | H | I | W | X | Y | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Actual | | | | | | | | | | | | |
| A | 1272 | 0 | 0 | 6 | 6 | 2 | 0 | 0 | 0 | 0 | 0 | 1286 |
| B | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| C | 6 | 24 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 31 |
| D | 1 | 0 | 104 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 106 |
| E | 16 | 0 | 4 | 302 | 3 | 0 | 0 | 0 | 0 | 3 | 0 | 328 |
| F | 45 | 0 | 0 | 0 | 537 | 5 | 1 | 0 | 0 | 0 | 0 | 588 |
| G | 7 | 0 | 0 | 0 | 0 | 127 | 0 | 0 | 0 | 0 | 0 | 134 |
| H | 6 | 1 | 0 | 0 | 2 | 0 | 147 | 0 | 0 | 0 | 0 | 156 |
| I | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 248 | 0 | 0 | 0 | 249 |
| W | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 13 |
| X | 3 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 148 | 2 | 157 |
| Y | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 75 | 80 |
| All | 1360 | 25 | 108 | 315 | 548 | 135 | 149 | 248 | 11 | 153 | 77 | 3129 |

```
Features sorted by their score:
[(0.1928, 'intercolumnar_distance'), (0.1794, 'row_number'), (0.1786, 'upper_margin'), (0.1491, 'exploitation'), (0.1452, 'lower_margin'), (0.0662,
'peak_number'), (0.0504, 'interlinear_spacing'), (0.0235, 'modular_ratio'), (0.0148, 'weight')]
```

# Scatterplot of Most Used Features (Axes are Limited)