Predicting Probability of Arrest in Cases of Use of Force by Police

Andrew Gibbs-Bravo | 180028746 | INM430

Notebook HTML Link: https://smcse.city.ac.uk/student/aczd071/AGB-1.html

Word count (incl. titles): Analytical Question: 518 | Findings: 1,054

Analytical Question and Approach

Overview of Question and Domain

The analytical question being considered is **"Can one predict whether a subject was arrested in instances where police applied force"**. The main dataset used in the analysis is from the London Datastore and consists of ~270 features describing ~90,000 instances of police officers using force on subjects between April 2017 (when dataset what introduced) and August 2018. [1] The dataset states whether the subject was arrested, therefore **the objective is to create a binary classifier with the highest test arrest prediction accuracy**.

When originally exploring the dataset, I assumed that the vast majority of instances would result in arrest although only ~66% resulted in arrest with other outcomes including: hospitalization, the subject escaping, detention under mental health act, and the subject simply not being arrested.

One of the main technical challenges of this dataset is that in addition to the relatively high number of features, (compared to beginner Kaggle datasets) the vast majority of features are categorical variables which makes feature engineering and dimensionality reduction challenging. A secondary analytical question was therefore **"What are the best approaches for extracting information from categorical features"**.

The original dataset was also supplemented with another dataset from the London Datastore which records offense rates by type within each borough. [2]

Analytical Approach

Initial exploration, cleaning, and pre-processing

- Merged two years of Use of Force data and included borough criminal and violent offence rate datasets
- Examined each feature's datatype, sparsity, type of statistic, relevant summary statistics, and created an issues list
- Removed duplicate observations in instances which shared the same date, borough, location, subject ethnicity, age and gender, and time of day
- Several features required manipulations and transformations

Exploratory data analysis, feature engineering, feature extraction and modelling

- Conducted visual exploratory analysis of the relationships in the data
    - Used Tableau for geographic based data exploration
    - Created a variety of visualizations in order to understand relationships between features and arrest rate
- Experimented with manual feature engineering and automated feature engineering (Featuretools library) although results were very poor
- Developed an embedded pipeline with different model configurations in order to be able to easily test the impact on performance. Some transformations examined in the pipeline include:
    - Changing the encoding of categorical features (one-hot vs mean-encoded)
    - Removing one of the correlated variables where two are highly correlated
    - Changing the dimensionality reduction technique applied including:

        - Principal Components Analysis (PCA), autoencoder, Latent Dirichlet Allocation (LDA), Multiple Components Analysis (MCA), Nonnegative Matrix Factorization (NMF)
    - Applying feature selection (lasso regularization)
    - Adding K-Modes clusters as a feature
- Applied models on cross validation set to determine optimal configuration. Applied:
    - Logistic regression: a linear classification model
    - LightGBM: a nonlinear classification model which applies gradient boosted decision trees
- Tuned model hyperparameters using full training set with optimal configuration and 3-fold cross validation gridsearch
- Analyzed performance of models on holdout test set in terms of accuracy while also considering AUC, precision, recall, and f-1 score
- Examined feature importance of features in final model configuration from random forest

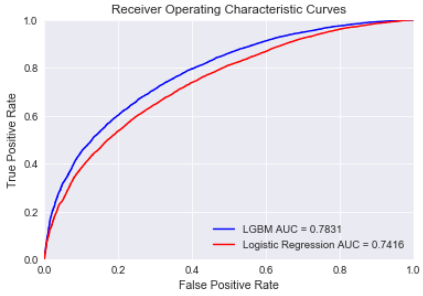
Summary of Findings

Significant Findings

Models are Able to Improve Prediction Accuracy

In predicting whether the subject would be arrested given a use of force by police, you would be able to achieve ~66% accuracy by predicting that they would be arrested. If you were to use the best performing model, you would be able to increase the prediction accuracy by 12% to 74%.

LightGBM (LGBM) consistently performs better than logistic regression (LR) which is as expected



Light GBM

(Accuracy: 0.7412)

		Prediction	
		No Default	Default
Actual	No Default	4106	4918
	Default	2057	15866

AUC	0.7831
Precision	0.7634
Recall	0.8852
F-1 Test	0.8198

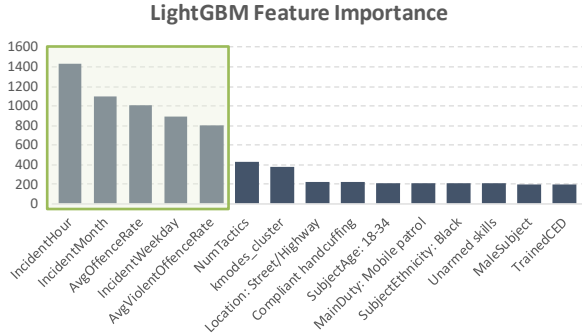
Logistic Regression

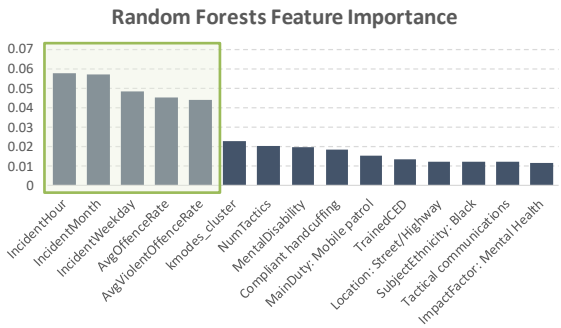
(Accuracy: 0.7143)

		Prediction	
		No Default	Default
Actual	No Default	3271	5753
	Default	1947	15976

AUC	0.7416
Precision	0.7352
Recall	0.8914
F-1 Test	0.8058

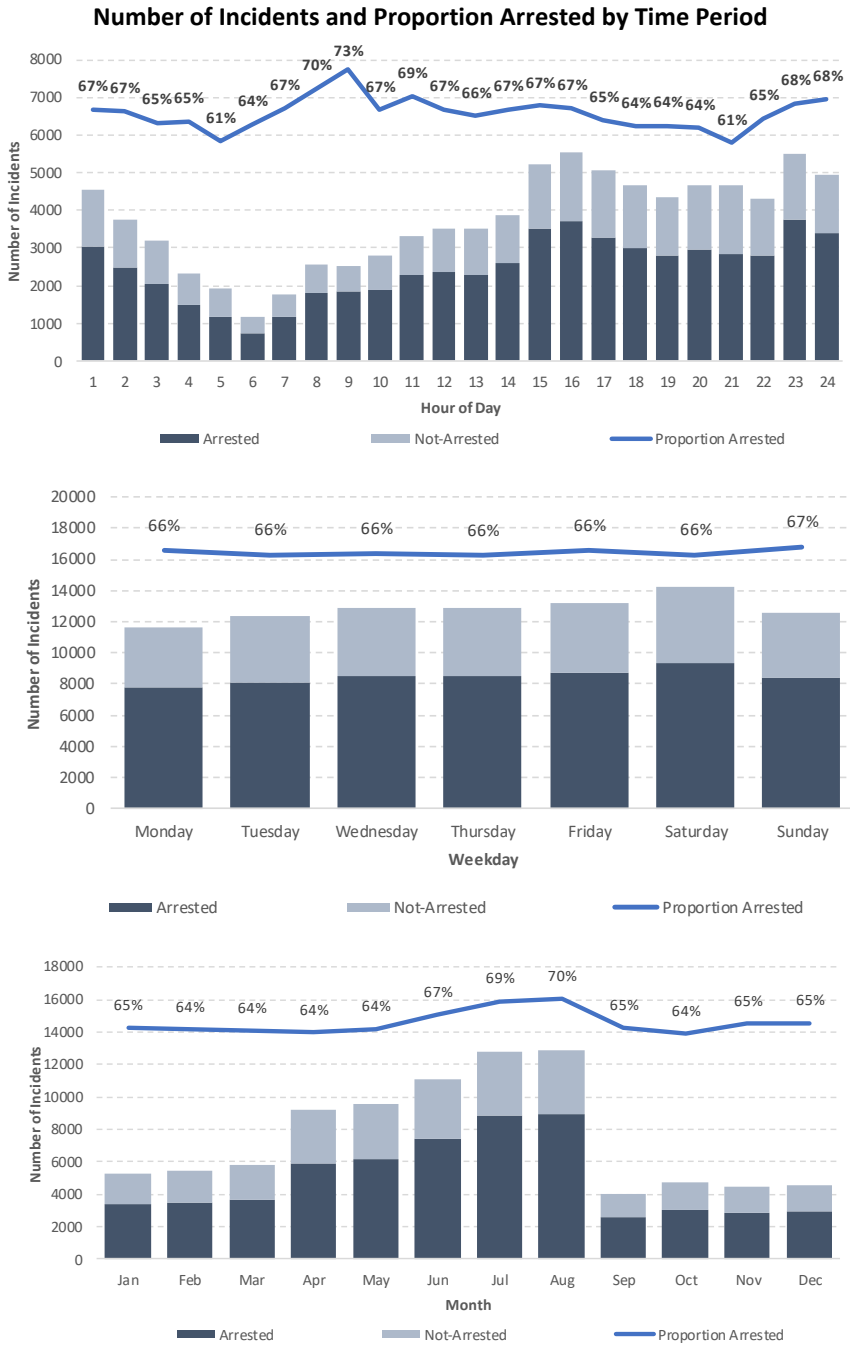
The features which were the most important based on LGBM and checked against random forests feature importance were features related to time, location, and then demographic information / details of the interaction.





Time Based Features

Time based features are surprisingly important to the prediction. There is a significant amount of variation in the arrest rates on an hourly and monthly basis, although minimal variation by weekday.



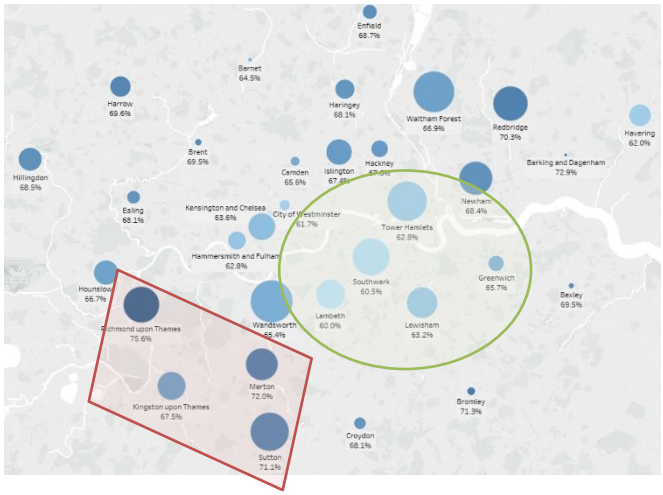
- *Note: Monthly data has not been adjusted and therefore only considers data from April 2017 to August 2018*

Location Based Features

Borough Location

Arrest rates vary considerably depending on borough with clusters of boroughs with lower arrest rates and others with higher arrest rates. For example, the area highlighted in green has a lower proportion of arrests given a comparable number of incidents (size of circles) versus the area highlighted in red.

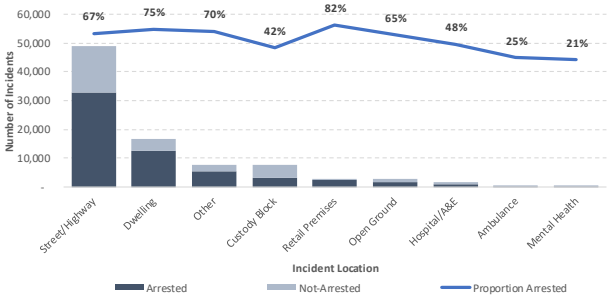
Number of Incidents and Proportion Arrested by Borough



Incident Location

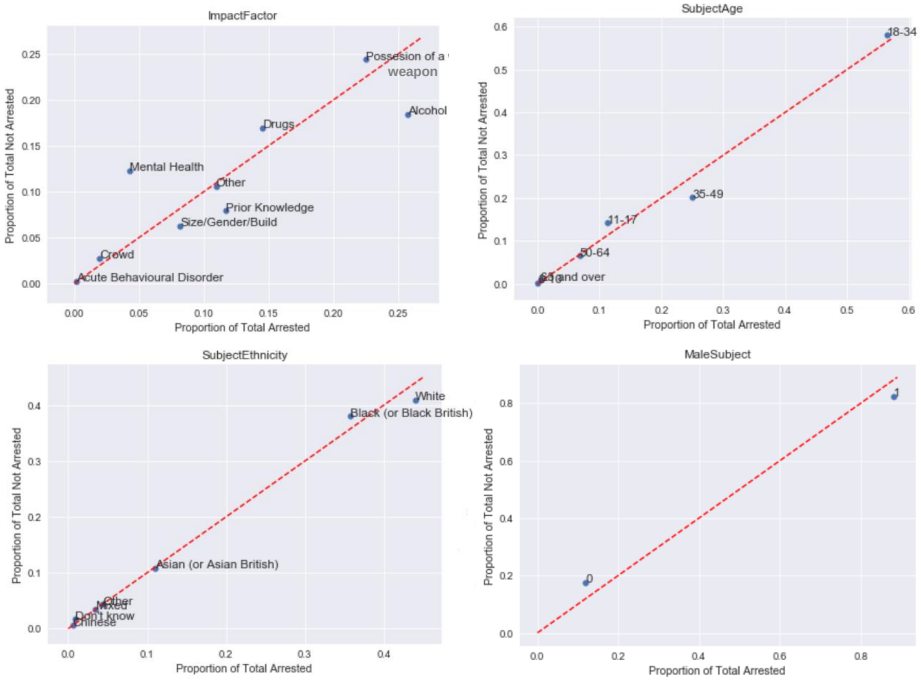
The majority of incidents occur on the street / highway, however, it is evident that those which occur in dwellings or retail premises have much higher arrest rates likely driven by the reason the officer was involved. Incidents which occur in medical and mental health situations have much lower arrest rates.

Number of Incidents and Proportion Arrested by Incident Location



Demographic Based Features

These charts represent the proportion of the total arrested and not arrested with a given characteristic. We can conclude that alcohol, being between 35-49 years old, and male results in a higher likelihood of being arrested. However, please refer to the limitations section for a discussion around the challenges in interpretation.

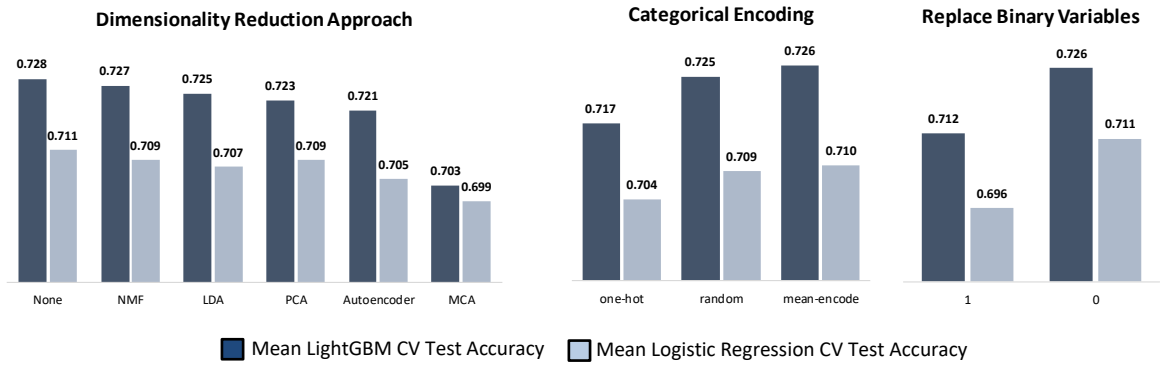


Categorical Variables are Challenging to Extract Novel Information From

The secondary question in the report was what approaches are most useful for extracting information from categorical features and unfortunately the analysis was largely unsuccessful.

Significant findings from this dataset relating to modelling include:

- The difference in accuracy is not very substantial between approaches
    - The top performing configuration achieved test accuracy of 0.7325 versus the 50th ranked approach of 0.7264
- Dimensionality reduction approaches resulted in underwhelming performance with NMF and LDA tending to perform somewhat better
- Replacing one-hot encoded data with dimensionality reduced components resulted in poor performance
- K-modes was slightly additive to model performance
    - Top five configurations all include K-folds
- Performance stability increases when increasing CV and test split
    - Prior version had test split and CV split of only 0.15 which led to instability in model selection which improved significantly when increasing split to 0.30
- Model performance increased when retraining on full dataset with hyperparameter tuning
    - Final model achieved test accuracy of 0.7412 with LGBM versus highest CV of 0.7325



The correlation of the results between LGBM and LR is high across model configurations for AUC at 0.73 and for the test accuracy 0.82 although it is almost zero between training sets 0.03 for a given hyperparameter setting. This is due to LGBM having negative correlation -0.65 between training and test performance for a given hyperparameter configuration as the model tends to overfit and logistic regression having a high correlation of 0.90.

Application of Findings

Understanding Drivers of Prediction Accuracy of Subject Being Arrested

The weights and feature importance from the model can be used by police watchdog organizations such as the Independent Office for Police Conduct to ensure that certain groups aren't being disproportionately arrested or (using supplemental data) ensure that police aren't using force without reason.

Applying machine learning models to justice is an evolving practice although tools such as COMPAS are already being applied in the United States legal system in bail hearings and sentencing. [3,4,5] The relatively poor performance in this model highlights the challenges of applying machine learning models prescriptively to social problems.

The most informative features in the best performing model relate to time of incident and borough which is likely problematic unless there is a confounding variable of arrestable crimes increasing during these times in these areas. Therefore, I would not propose this model be used by police in deciding whether a subject is arrested or for commercial purposes.

Proposing an Approach for Extracting Information from Categorical Variables

Some findings which are generalizable to other situations when dealing with large scale categorical features are:

- Embedded approach works although does not scale given computational resources required
- Focusing on feature engineering, understanding the data, and supplementing the data rather than tweaking model structure is likely to result in far greater improvements in performance
- The performance did not vary significantly between dimensionality reduction techniques, therefore should first attempt simple approaches and then test one or two dimensionality reduction approaches
- The high correlation between algorithm test results demonstrates if you only had the resources to train one model you should train LGBM and can expect positive changes to generalize to logistic regression

Potential Sources of Bias and Limitations

There are several potential sources of bias in the dataset which result in limitations in its interpretation:

- **Selective / inaccurate reporting:** Instances are reported by officers after they occur which can potentially result in selective reporting (omitting instances) and an inconsistent recollection of events.
- **Insufficient context in the dataset:** The dataset does not contain reason for incident which is presumed to be significantly predictive and is a confounding variable of the analysis. Refer to future work.
- **Limitations in using only one source of data for prediction**
    - For example, given white subjects are more likely to be arrested given a use of force than black subjects it can be concluded that there is a bias towards arresting white subjects.
    - Without knowing the context for each incident, it can also be concluded that black subjects are more likely to have force against them without an arrestable offense.

Appendix

Sources

1. Use of force dataset: https://data.london.gov.uk/dataset/use-of-force
2. Recorded offense dataset: https://data.london.gov.uk/dataset/recorded_crime_rates
3. https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/
4. https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/
5. https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html
6. https://www.theguardian.com/uk-news/2018/oct/08/met-polices-use-of-force-jumps-79-in-one-year

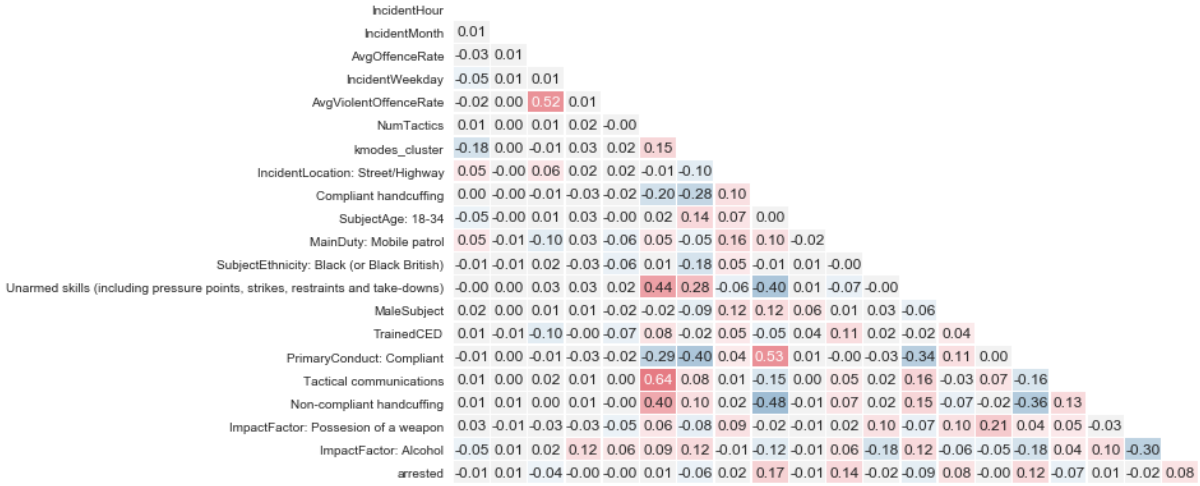
Future Work

Given the time restriction there were many approaches that may improve the results although I was unable to which could be completed in future work:

- Optimize hyperparameters for dimensionality reduction techniques (e.g. select optimal number of components for PCA)
- Add additional dimensionality reduction techniques designed for categorical variables
- Run an exhaustive search on more training data. Alternatively, search model configuration search space more intelligently than a random search such as using evolutionary algorithms
- Can try additional algorithms to LR and LGBM
- Can do additional feature engineering
- Can parallelize gridsearch and leverage a GPU for some algorithms to boost speed
- Supplement dataset to consider relationship with other crime statistics and census data to analyze the relationship between use of force relative to general population and demographics. The Guardian [6] provides high level analysis of this nature although their work is deficient in that it only compares to the general population and does not consider the crime rates of the populations.

Correlation Heatmap of the Most Informative Features

The minimal correlation between the target and most of the highly predictive variables means there are non-linear relationships between the variables which explains why LGBM does significantly better than LR



Random Sample of Model Configuration Performance

Config	sel_cat_encoding	feature_selection	remove_corr	lgbm_AUC	lgbm_train_acc	lgbm_test_acc	logr_AUC	logr_train_acc	logr_test_acc	cat_encoding	binary_dim_reduce	replace_binary_vals	add_kmodes
1	random	FALSE	FALSE	0.7683	0.7681	0.7325	0.7328	0.7122	0.7097	-	None	FALSE	TRUE
2	one-hot	FALSE	FALSE	0.7699	0.7681	0.7314	0.7350	0.7152	0.7145	-	None	FALSE	TRUE
3	random	TRUE	FALSE	0.7673	0.7809	0.7304	0.7348	0.7159	0.7127		LDA	FALSE	TRUE
4	mean-encode	TRUE	TRUE	0.7662	0.7713	0.7303	0.7214	0.7132	0.7105	-	LDA	FALSE	TRUE
5	random	TRUE	FALSE	0.7659	0.7732	0.7302	0.7333	0.7158	0.7143		NMF	FALSE	TRUE
6	random	TRUE	TRUE	0.7667	0.7721	0.7299	0.7186	0.7084	0.7079		NMF	TRUE	FALSE
7	mean-encode	FALSE	FALSE	0.7657	0.7698	0.7293	0.7218	0.7134	0.7097	-	None	FALSE	TRUE
8	mean-encode	TRUE	TRUE	0.7678	0.7697	0.7292	0.7203	0.7127	0.7107	-	NMF	FALSE	FALSE
9	one-hot	FALSE	FALSE	0.7682	0.7637	0.7292	0.7349	0.7149	0.7132	-	None	FALSE	FALSE
10	mean-encode	FALSE	FALSE	0.7665	0.7708	0.7291	0.7225	0.7130	0.7102	-	NMF	FALSE	FALSE
11	mean-encode	FALSE	FALSE	0.7665	0.7708	0.7291	0.7225	0.7130	0.7102	-	NMF	FALSE	FALSE
12	random	FALSE	FALSE	0.7660	0.7666	0.7290	0.7323	0.7139	0.7113		None	FALSE	FALSE
13	one-hot	TRUE	TRUE	0.7656	0.7643	0.7290	0.7360	0.7151	0.7118	-	None	FALSE	TRUE
14	mean-encode	FALSE	TRUE	0.7670	0.7686	0.7289	0.7214	0.7132	0.7105	-	LDA	FALSE	FALSE
15	mean-encode	TRUE	TRUE	0.7661	0.7679	0.7288	0.7203	0.7133	0.7096	-	Autoencoder	FALSE	FALSE
16	random	TRUE	TRUE	0.7672	0.7666	0.7287	0.7293	0.7157	0.7128		None	FALSE	TRUE
17	one-hot	TRUE	TRUE	0.7678	0.7635	0.7284	0.7349	0.7148	0.7124	-	None	FALSE	FALSE
18	random	TRUE	TRUE	0.7672	0.7681	0.7283	0.7243	0.7139	0.7092		None	FALSE	FALSE
19	mean-encode	TRUE	FALSE	0.7678	0.7676	0.7283	0.7227	0.7134	0.7101	-	LDA	FALSE	FALSE
20	mean-encode	FALSE	FALSE	0.7680	0.7692	0.7283	0.7224	0.7131	0.7105	-	Autoencoder	FALSE	FALSE
21	random	TRUE	FALSE	0.7661	0.7660	0.7282	0.7279	0.7136	0.7112		None	FALSE	FALSE
22	mean-encode	TRUE	TRUE	0.7668	0.7703	0.7281	0.7201	0.7134	0.7105	-	NMF	FALSE	TRUE
23	mean-encode	TRUE	FALSE	0.7656	0.7671	0.7281	0.7228	0.7132	0.7095	-	NMF	FALSE	TRUE
24	mean-encode	FALSE	TRUE	0.7666	0.7712	0.7280	0.7196	0.7133	0.7108	-	None	FALSE	TRUE
25	mean-encode	FALSE	FALSE	0.7662	0.7696	0.7280	0.7237	0.7136	0.7093	-	LDA	FALSE	TRUE
26	mean-encode	FALSE	TRUE	0.7655	0.7704	0.7280	0.7199	0.7131	0.7104	-	Autoencoder	FALSE	TRUE
27	mean-encode	TRUE	TRUE	0.7666	0.7687	0.7277	0.7243	0.7138	0.7121	-	LDA	FALSE	TRUE
28	one-hot	TRUE	FALSE	0.7668	0.7650	0.7277	0.7350	0.7145	0.7120	-	NMF	FALSE	FALSE
29	mean-encode	TRUE	TRUE	0.7660	0.7701	0.7275	0.7196	0.7132	0.7106	-	None	FALSE	FALSE
30	one-hot	TRUE	FALSE	0.7631	0.7798	0.7275	0.7347	0.7142	0.7134	-	Autoencoder	FALSE	FALSE
31	mean-encode	TRUE	TRUE	0.7628	0.7672	0.7274	0.7191	0.7132	0.7093	-	None	FALSE	TRUE
32	random	FALSE	FALSE	0.7614	0.7816	0.7273	0.7250	0.7161	0.7125		LDA	FALSE	TRUE
33	mean-encode	FALSE	TRUE	0.7665	0.7684	0.7273	0.7196	0.7133	0.7104	-	None	FALSE	FALSE
34	mean-encode	TRUE	FALSE	0.7665	0.7702	0.7273	0.7222	0.7131	0.7102	-	PCA	FALSE	FALSE
35	one-hot	FALSE	FALSE	0.7597	0.7869	0.7273	0.7363	0.7144	0.7131	-	PCA	FALSE	TRUE
36	mean-encode	FALSE	FALSE	0.7662	0.7694	0.7272	0.7218	0.7130	0.7109	-	Autoencoder	FALSE	TRUE
37	one-hot	TRUE	TRUE	0.7667	0.7637	0.7271	0.7346	0.7147	0.7122	-	None	FALSE	TRUE
38	mean-encode	FALSE	TRUE	0.7660	0.7701	0.7270	0.7199	0.7120	0.7093	-	Autoencoder	FALSE	FALSE
39	random	TRUE	TRUE	0.7663	0.7687	0.7270	0.7266	0.7150	0.7118		None	FALSE	TRUE
40	one-hot	TRUE	TRUE	0.7635	0.7806	0.7269	0.7371	0.7148	0.7101	-	NMF	FALSE	TRUE
41	mean-encode	FALSE	FALSE	0.7655	0.7679	0.7267	0.7231	0.7129	0.7105	-	LDA	FALSE	TRUE
42	one-hot	TRUE	FALSE	0.7604	0.7834	0.7267	0.7382	0.7175	0.7125	-	LDA	FALSE	TRUE
43	mean-encode	TRUE	FALSE	0.7693	0.7703	0.7265	0.7219	0.7122	0.7109	-	Autoencoder	FALSE	FALSE
44	mean-encode	FALSE	TRUE	0.7677	0.7714	0.7265	0.7203	0.7128	0.7104	-	NMF	FALSE	FALSE
45	mean-encode	FALSE	TRUE	0.7677	0.7714	0.7265	0.7203	0.7128	0.7104	-	NMF	FALSE	FALSE
46	one-hot	FALSE	FALSE	0.7621	0.7859	0.7265	0.7355	0.7145	0.7127	-	Autoencoder	FALSE	FALSE
47	one-hot	TRUE	TRUE	0.7615	0.7824	0.7265	0.7360	0.7170	0.7110	-	LDA	FALSE	FALSE
48	random	FALSE	FALSE	0.7633	0.7834	0.7265	0.7284	0.7160	0.7113		LDA	FALSE	FALSE
49	mean-encode	TRUE	TRUE	0.7632	0.7678	0.7265	0.7184	0.7117	0.7080	-	Autoencoder	FALSE	TRUE
50	random	TRUE	TRUE	0.7647	0.7833	0.7264	0.7307	0.7140	0.7126		PCA	FALSE	TRUE
51	random	FALSE	FALSE	0.7631	0.7810	0.7264	0.7312	0.7148	0.7110		NMF	FALSE	FALSE
52	random	FALSE	FALSE	0.7600	0.7771	0.7264	0.7114	0.7063	0.7089		PCA	TRUE	FALSE
53	one-hot	FALSE	TRUE	0.7666	0.7662	0.7261	0.7346	0.7148	0.7129	-	None	FALSE	FALSE
54	one-hot	FALSE	FALSE	0.7607	0.7846	0.7261	0.7352	0.7145	0.7125	-	PCA	FALSE	FALSE
55	random	FALSE	FALSE	0.7623	0.7795	0.7259	0.7276	0.7147	0.7108		PCA	FALSE	FALSE
56	random	TRUE	FALSE	0.7656	0.7825	0.7257	0.7293	0.7161	0.7110		Autoencoder	FALSE	TRUE
57	random	FALSE	TRUE	0.7626	0.7823	0.7257	0.7185	0.7046	0.7042		NMF	TRUE	TRUE
58	random	TRUE	FALSE	0.7573	0.7790	0.7255	0.7171	0.7078	0.7042		NMF	TRUE	TRUE
59	random	TRUE	TRUE	0.7628	0.7798	0.7255	0.7310	0.7152	0.7113		LDA	FALSE	TRUE
60	one-hot	TRUE	TRUE	0.7598	0.7818	0.7252	0.7365	0.7175	0.7144	-	LDA	FALSE	TRUE
61	random	TRUE	TRUE	0.7621	0.7793	0.7250	0.7314	0.7136	0.7111		Autoencoder	FALSE	TRUE
62	one-hot	TRUE	TRUE	0.7619	0.7868	0.7250	0.7371	0.7153	0.7121	-	NMF	FALSE	FALSE
63	random	TRUE	FALSE	0.7614	0.7752	0.7250	0.7285	0.7145	0.7116		Autoencoder	FALSE	FALSE
64	one-hot	FALSE	FALSE	0.7596	0.7856	0.7249	0.7387	0.7171	0.7134	-	LDA	FALSE	TRUE
65	one-hot	FALSE	FALSE	0.7599	0.7863	0.7249	0.7363	0.7151	0.7131	-	Autoencoder	FALSE	TRUE
66	mean-encode	FALSE	TRUE	0.7651	0.7700	0.7247	0.7201	0.7131	0.7105	-	PCA	FALSE	FALSE
67	mean-encode	TRUE	TRUE	0.7654	0.7695	0.7244	0.7200	0.7130	0.7106	-	PCA	FALSE	FALSE
68	random	TRUE	FALSE	0.7611	0.7833	0.7241	0.7290	0.7162	0.7127		LDA	FALSE	FALSE
69	one-hot	FALSE	FALSE	0.7608	0.7846	0.7237	0.7384	0.7168	0.7125	-	LDA	FALSE	FALSE
70	one-hot	TRUE	FALSE	0.7590	0.7813	0.7234	0.7352	0.7141	0.7126	-	PCA	FALSE	FALSE
71	random	TRUE	TRUE	0.7496	0.7776	0.7234	0.6977	0.6972	0.6960		LDA	TRUE	FALSE
72	random	FALSE	TRUE	0.7634	0.7695	0.7234	0.7299	0.7135	0.7111		None	FALSE	TRUE
73	random	TRUE	TRUE	0.7572	0.7780	0.7231	0.7194	0.7105	0.7091		PCA	TRUE	TRUE
74	one-hot	TRUE	TRUE	0.7582	0.7835	0.7229	0.7349	0.7143	0.7125	-	PCA	FALSE	FALSE
75	random	FALSE	TRUE	0.7561	0.7769	0.7225	0.7158	0.7094	0.7068		PCA	TRUE	FALSE
76	mean-encode	TRUE	TRUE	0.7605	0.8065	0.7224	0.7284	0.7148	0.7074	-	MCA	FALSE	FALSE
77	random	FALSE	TRUE	0.7605	0.7801	0.7222	0.7311	0.7130	0.7115		Autoencoder	FALSE	TRUE
78	one-hot	FALSE	FALSE	0.7613	0.7884	0.7221	0.7372	0.7149	0.7117	-	NMF	FALSE	FALSE
79	random	TRUE	FALSE	0.7573	0.7748	0.7217	0.7177	0.7094	0.7090		PCA	TRUE	FALSE
80	mean-encode	FALSE	FALSE	0.7579	0.8161	0.7212	0.7335	0.7161	0.7053	-	MCA	FALSE	TRUE
81	mean-encode	FALSE	TRUE	0.7599	0.8135	0.7199	0.7280	0.7149	0.7087	-	MCA	FALSE	TRUE
82	random	TRUE	TRUE	0.7569	0.7729	0.7199	0.7178	0.7109	0.7097		PCA	TRUE	FALSE
83	mean-encode	FALSE	FALSE	0.7598	0.8116	0.7186	0.7323	0.7153	0.7048	-	MCA	FALSE	TRUE
84	mean-encode	TRUE	FALSE	0.7588	0.8070	0.7176	0.7320	0.7151	0.7055	-	MCA	FALSE	FALSE
85	random	FALSE	FALSE	0.7468	0.7748	0.7167	0.7055	0.6999	0.6977		Autoencoder	TRUE	TRUE
86	random	FALSE	FALSE	0.7394	0.7737	0.7146	0.6800	0.6994	0.6927		LDA	TRUE	FALSE
87	one-hot	TRUE	FALSE	0.7419	0.7848	0.7146	0.6956	0.6935	0.6943	-	NMF	TRUE	TRUE
88	one-hot	TRUE	TRUE	0.7380	0.7776	0.7136	0.7009	0.6961	0.6973	-	PCA	TRUE	TRUE
89	one-hot	FALSE	TRUE	0.7311	0.7741	0.7123	0.6755	0.6893	0.6887	-	LDA	TRUE	TRUE
90	one-hot	FALSE	TRUE	0.7380	0.7760	0.7122	0.7010	0.6964	0.6976	-	PCA	TRUE	FALSE
91	one-hot	FALSE	FALSE	0.7323	0.7790	0.7121	0.6613	0.6852	0.6819	-	LDA	TRUE	TRUE
92	one-hot	TRUE	TRUE	0.7261	0.7721	0.7090	0.6785	0.6913	0.6852	-	Autoencoder	TRUE	TRUE
93	one-hot	TRUE	TRUE	0.7407	0.8408	0.7087	0.7348	0.7129	0.7031	-	MCA	FALSE	TRUE
94	one-hot	FALSE	TRUE	0.7281	0.7694	0.7082	0.6811	0.6940	0.6936	-	Autoencoder	TRUE	FALSE
95	random	FALSE	TRUE	0.7271	0.8527	0.7002	0.7228	0.7177	0.6957		MCA	FALSE	FALSE
96	one-hot	FALSE	FALSE	0.7076	0.7676	0.6978	0.6448	0.6807	0.6779	-	Autoencoder	TRUE	TRUE
97	one-hot	TRUE	FALSE	0.7048	0.7673	0.6941	0.6432	0.6812	0.6781	-	Autoencoder	TRUE	FALSE
98	one-hot	FALSE	FALSE	0.7298	0.8541	0.6913	0.7274	0.7142	0.6961	-	MCA	FALSE	TRUE
99	one-hot	FALSE	FALSE	0.6970	0.8478	0.6718	0.6972	0.6978	0.6828	-	MCA	TRUE	TRUE
100	one-hot	TRUE	TRUE	0.6879	0.8313	0.6588	0.6975	0.6971	0.6832	-	MCA	TRUE	FALSE

Note: MCA was trained on only 25% of the total data given the long running time and poor initial results. When previously trained on larger proportions MCA still performed the worst out of the examined algorithms

MCA Visualization of Binary Features

The color shows the 5 k-mode clusters on two dimensions of MCA which helps to visualize why the kmode feature may be useful in prediction. Note that this accounts for under ~20% of variance in the data and therefore the visualization should not have too much weight placed on it.