WORLD OF BARGAINS

STORE PERFORMANCE

ASSIGNMENT REPORT

Student Number: 2938740

ITNPBD6 Data Analytics

Spring 2021

CONTENTS

Contents 2

1.INTRODUCTION 4

1.1.TASK 4

1.2.APPROACH 5

1.3.GLOSSARY OF TERMS 5

2.DATA UNDERSTANDING 6

2.1.DATA SUMMARY 6

2.2.MISSING VALUES & DATA ERRORS 7

2.3.DATA DISTRIBUTIONS 7

2.4.CORRELATED DATA 10

3.DATA PREPARATION 11

3.1.SPLITTING THE DATA 11

3.2.CLEANING THE DATA 11

3.3.FEATURE ENGINEERING 12

3.4.DATA SCALING 12

3.5.OUTLIER DETECTION 12

4.MODELS 13

4.1.EQUALITY OF OUTCOME 13

4.2.RANDOM FOREST DECISION TREE 13

4.2.1.BASELINE MODEL 13

4.2.2.HYPERPARAMETER TUNING 13

4.2.3.FEATURE SELECTION 14

4.2.4.CONCLUSION 14

4.3.LOGISTIC REGRESSION CLASSIFIER 15

4.3.1.BASELINE MODEL 15

4.3.2.HYPERPARAMETER TUNING 15

4.3.3.FEATURE SELECTION 16

4.3.4.CONCLUSION 16

4.4.MULTI-LAYER PERCEPTRON (NEURAL NETWORK) 17

4.4.1.BASELINE MODEL 17

4.4.2.HYPERPARAMETER TUNING 17

4.4.3.FEATURE SELECTION 18

4.4.4.CONCLUSION 18

4.5.FINAL MODEL 19

5.EVALUATION 20

5.1.TEST SET RESULTS 20

5.2.CONFUSION MATRIX 20

5.3.AUC & ROC 21

5.4.MODEL ERRORS 21

5.5.DEPLOYMENT & SUMMARY OF CRISP-DM 22

1. INTRODUCTION

1.1.TASK

With over 100 stores in the UK, the client (World of Bargains) is looking to expand the business efficiently and effectively. It is important for the client have a documented scientific method of predicting which store features will likely end with a profitable investment and achieve the goal of growth whilst minimising the losses associated with opening unprofitable stores.

To facilitate the task the client has provided a subset of data from their existing stores along with a simple classification metric of performance, which is described as either 'Good' or 'Bad'. This makes the project a binary classification task and will require appropriate models to be developed. The absence of any financial data means that the predictions will be made from store attributes and is a clear sign that the client is interested in determining whether it is possible to predict a store's profitability from non-financial data.

Although the client has not defined the requirements of what constitutes a good model it is reasonable to define two metrics that will be important in assessing the quality of the models that will be developed for evaluation:

- Accuracy: The model must at least do better than a baseline of guessing whether a new store will be profitable.
- Precision: The model will need to minimise False Positives as the cost of opening an unprofitable store is at odds with the stated goal of effective company growth. This will be the key evaluation metric.
- F1 & Recall: These metrics will also be provided during the selection process to help separate identical models.

The model with the best accuracy will not automatically be chosen if it has a poor precision rate as this would leave the client at serious risk of investing in ultimately unprofitable stores that could severely damage or even bankrupt the client in the worst-case scenario. Conversely, False Negatives pose no financial risk to the client if the stores end up being unexpectedly profitable.

1.2.APPROACH

A data mining approach to this task will be suitable as there is no clearly identifiable relationship in the data between the stores' attributes and their profitability. The absence of any sales and revenue data is well noted which eliminates the prospect of generating a quantitative prediction through regression analysis. The development of machine learning models will hopefully identify patterns in the data that cannot be seen by the human eye.

The CRISP-DM methodology will be used to provide a structured approach to the analysis and model development with the 6 key stages of this methodology forming the section headers used in this document.

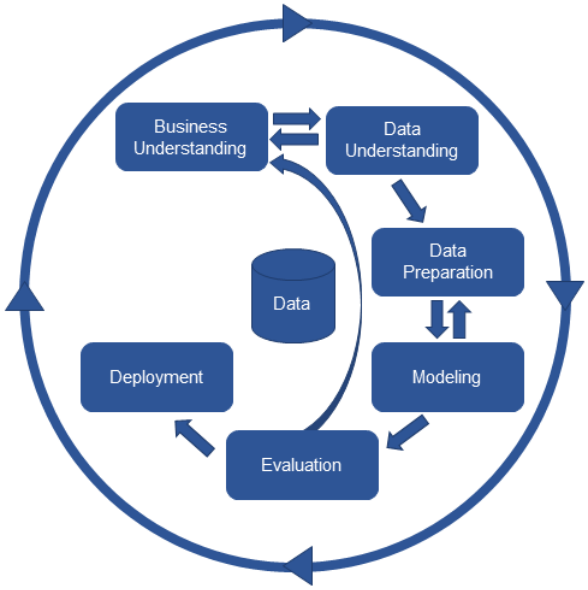

Figure 1: Stages of the CRISP-DM Model

1.3.GLOSSARY OF TERMS

Confusion over terminology is something we wish to avoid therefore a brief glossary of key terms is given:

Term	Definition
Model	A term used to describe the computer generated mathematical model of the data
Variable / Feature	The data associated with each column in the dataset, made up of observations (rows). The two terms are used interchangeably
Target Feature	The target feature is the feature that we are trying to predict, in this case, performance.
CRISP-DM	A framework used in data driven projects to give structure to the modelling process
Decision Trees	A connected network of filters that generate a decision based on a series of specific criteria
Logistic Regression	A mathematical technique used to estimate the parameters of predicting a binary outcome
Neural Network	A series of algorithms that attempts to identify patterns in data by mimicking the human brain

2. DATA UNDERSTANDING

2.1.DATA SUMMARY

A dataset was provided containing a subset of data from stores located in towns beginning with the letter 'S'. This implies that more data should be readily available if required. The dataset consists of 136 rows and 19 columns including the target variable of 'Performance' which is also well-balanced. No missing values were identified. A summary of the data is given below labelled as per data file:

Variable	Type	Role	Likely Useful?	Comments
Town	Text	Meta	**No**	Unique Values
Manager name	Text	Meta	**No**	Data is not useful as contains only first name
Country	Categorical (Nominal)	Feature	**No**	2 observations listed as 'France'. Company is UK. See section 2.2 for reasons to exclude.
Store ID	Numeric (Discrete)	Feature	**No**	Unique values
Staff	Numeric (Discrete)	Feature	Yes	Number of staff per store
Floor Space	Numeric (Continuous)	Feature	Yes	
Window	Numeric (Continuous)	Feature	Yes	
Car park **	Categorical (Nominal)	Feature	Yes	Needs cleaning (Yes, No, Y, N). Binary choice
Demographic score *	Numeric (Continuous)	Feature	Yes	How is score calculated? Raw or categorised?
Location	Categorical (Nominal)	Feature	Yes	(High Street, Retail Park, Shopping Centre, Village)
40min population	Numeric (Discrete)	Feature	Yes	Potential correlation issue for further investigation
30 min population	Numeric (Discrete)	Feature	Yes	Heavily skewed distribution
20 min population	Numeric (Discrete)	Feature	Yes	Heavily skewed distribution
10 min population	Numeric (Discrete)	Feature	Yes	Heavily skewed distribution
Store age	Numeric (Continuous)	Feature	Yes	
Clearance space	Numeric (Continuous)	Feature	Yes	
Competition number	Numeric (Discrete)	Feature	Yes	Count of stores
Competition score *	Numeric (Continuous)	Feature	Yes	How is score calculated? Raw or categorised?
Performance **	**Categorical (Nominal)**	**Target**	**Yes**	**(Bad, Good). Binary choice**

** The demographic and competition score fields are treated as continuous as it is unknown if the scores are taken from a list of pre-defined criterion or is a genuine score. To be safe, the generic definition has been used. This could be confirmed with the client.*

*** The car park and performance variables can also be described as 'Categorical (Binary)'*

As part of the CRISP-DM process it would be necessary to circle back to the client to establish how the 'Demographic score' and 'Competition score' features were created. The scores lie in a range between 10 and 19 inclusive which does not feel like a natural range for a score.

Throughout this section other queries arise which would also be addressed through discussion with the client.

2.2.MISSING VALUES & DATA ERRORS

No missing values were identified but a small number of data errors were identified. These will be fixed as stated below however in reality, a quick phone-call or email to the client would solve these errors with certainty very quickly:

Variable	Issue
Country	2 entries listed as 'France' but brief states the company operates stores in the UK. The entries will be amended to 'UK' during pre-processing.
Staff	Three data errors with values -2, 300 and 600 were spotted. The records will need to be excluded as we cannot expect to reasonably infer the actual value without contacting the client
Car park	Data contains 'Yes', 'Y', 'No' and 'N'. This field will be converted to a binary 0 and 1 field (No and Yes) during pre-processing.

Given that the brief states that the company operates in the UK with no mention of an international presence it is entirely reasonable to recode the 'France' country values as UK given that the town name associated with the observation is a valid UK town. After the data entries are converted from 'France' to 'UK' the feature becomes one with a single value and as such is therefore no longer useful for modelling.

2.3.DATA DISTRIBUTIONS

The dataset is well balanced in the target class 'Performance' and the dataset will not need rebalancing. The percentage split between 'Good' and 'Bad' performance is 51/49:

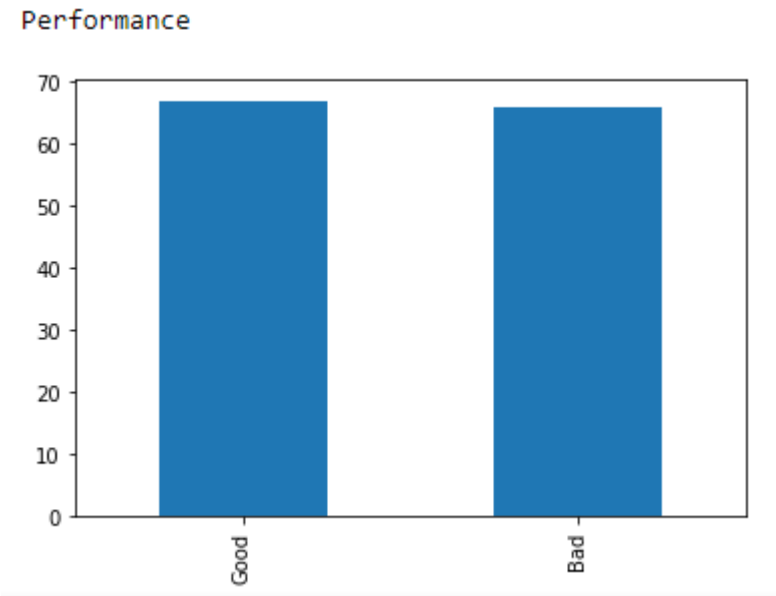

Figure 2: Chart showing the equal split of store performance in the dataset

As mentioned in the data summary table, the 'Town' and 'Store ID' features contained unique values so will be dropped. The 'Store ID' is self-explanatory as an ID field. The 'Town' distribution is evidenced here:

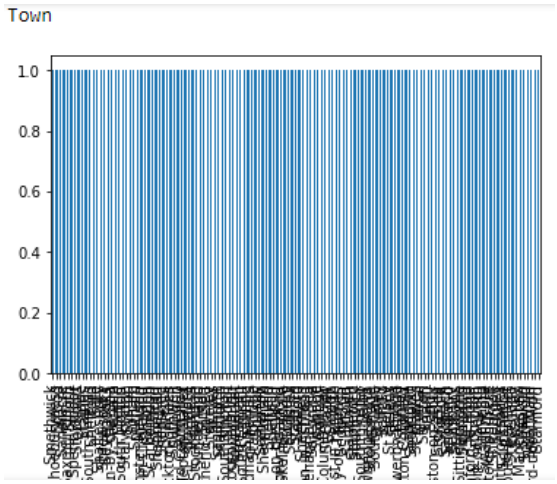

Figure 3: Chart showing homogeneity of Store ID field

The 'Manager name' feature is being dropped as it does not contain any meaningful data as shown by the distribution. At first glance it had potential to be useful as certain managers appeared responsible for multiple stores which gave rise to the possibility of a managerial skill feature however closer inspection revealed that the feature contained only first names and therefore it is possible that these are all different people who just happen to share a first name.

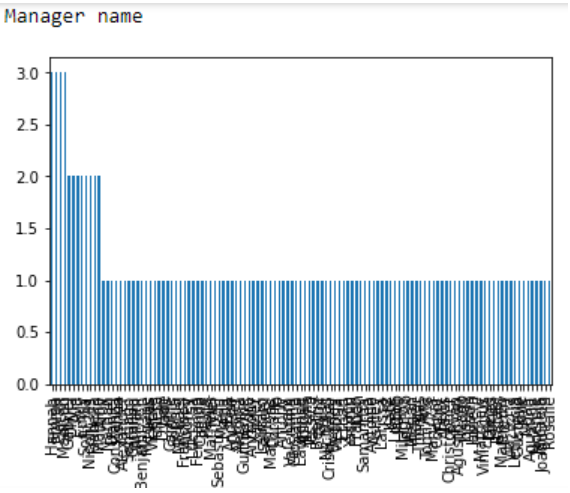

Figure 4: Chart showing distribution of manager name

If the client were able to provide data for the Manager ID then it may be possible to consider managerial skill as a valid feature for the model.

The breakdown of the population living within 10, 20, 30 and 40 minute drive shows very heavily skewed distributions for the 10, 20 and 30 minute populations. The '40min population' feature appears to be more evenly distributed (x-axis in millions):

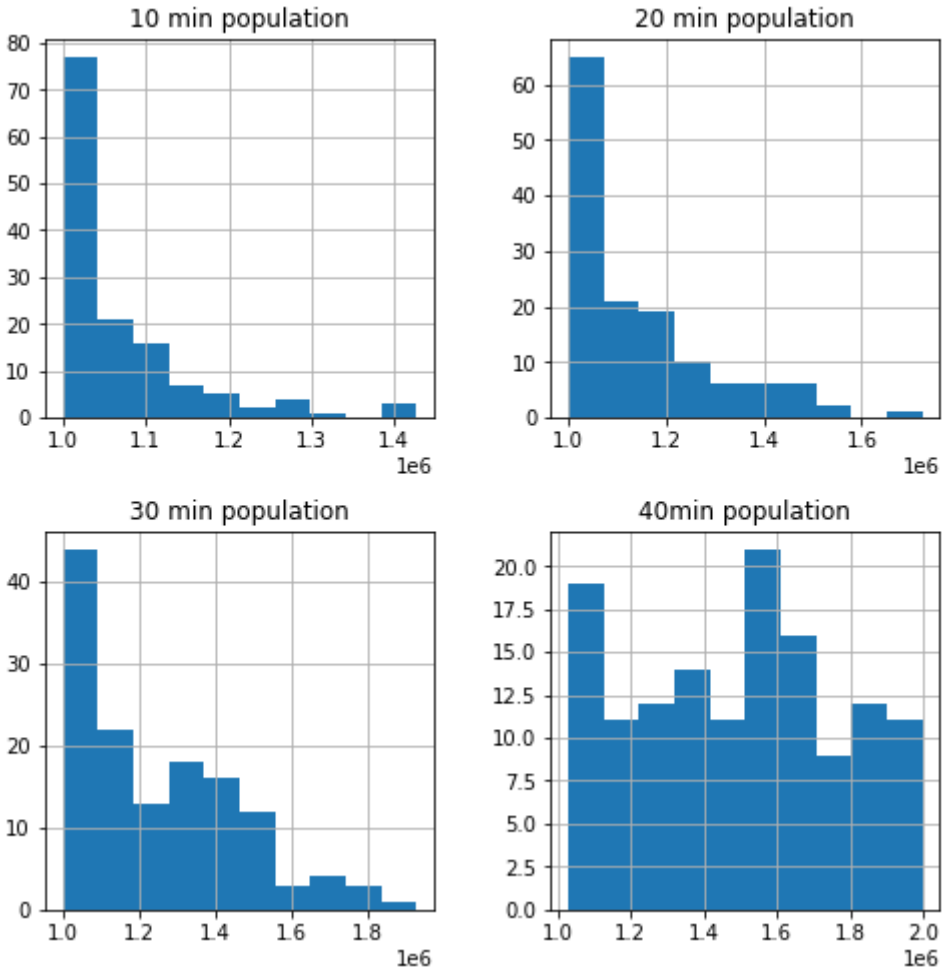

Figure 5: Chart showing distribution of population data

Without speaking to the client to gain a greater understanding of the lower boundary we cannot be certain that the lower boundary does not simply reflect an error in the data or that the populations have been manipulated to be just over the million mark. A quick check reveals that 107 of the 136 stores are located within a 10 minute drive of 1 – 1.1 million people. On the face of it, this seems statistically unlikely. More discussion is required to understand fully.

The remaining features appear to have a reasonably good distribution of data except for the 'Village' category within the 'Location' feature, which is very low. This may reflect that the client has tentatively entered some stores into a new market (literal village or outlet village). Once again, this would form part of the CRISP-DM loopback to gain a greater business understanding after reviewing the data.

2.4.CORRELATED DATA

A quick check of the data shows that there are some potential correlation issues that can be examined further with feature selection during the model generation process. There are two distinct clusters of features exhibiting varying degrees of correlation. The scores for both Pearson and Spearman correlation are given as the relationship between the data is difficult to see apart from the correlation between the Floor Space and Window features:

Cluster 1

Feature 1	Feature 2	Spearman	Pearson
Floor Space	Window	0.999	0.999
Floor Space	Clearance Space	0.663	0.627
Clearance Space	Window	0.662	0.629

Cluster 2

Feature 1	Feature 2	Spearman	Pearson
20 min population	10 min population	0.896	0.808
30 min population	20 min population	0.832	0.766
30 min population	10 min population	0.746	0.667
40min population	30 min population	0.597	0.625
40min population	20 min population	0.521	0.517
40min population	10 min population	0.464	0.433

A check for correlations against the target feature yielded no significant correlations:

Correlation Type	Feature Name	Correlation Value
Max	Staff	0.437251
Min	Location (High Street)	-0.335875

This is as far as we can go without splitting the data into train and test sets.

3. DATA PREPARATION

3.1.SPLITTING THE DATA

Before proceeding with the data preparation phase, it is important to split the dataset into test and train sets to avoid any form of data snooping bias and data leakage. The dataset will be split 80/20 with stratified sampling to preserve the balanced split in the target class. This is quite important given the small dataset size.

Dataset	Good Performance	Bad Performance
Training	55	53
Test	14	14
Total	**69**	**67**

The test set will be placed to one side and not used until the final model has been selected. K-Fold cross validation will be used to validate the training data at each step to preserve the integrity of the test data.

3.2.CLEANING THE DATA

There are several data cleansing actions that were previously identified and need to be completed before proceeding with the training set:

Step	Action	Comments
1	Clean Staff feature	Remove records with negative staff numbers and those greater than 10 The upper limit is arbitrary and would be discussed with the client in real-life This has been productionised to account for potential errors instead of exact dataset errors
2	Clean Parking feature	Set 'N' and 'No' to zero Set 'Y' and 'Yes' to 1
3	Convert Performance target	Set 'Bad' to zero Set 'Good' to 1
4	Encode Location feature (one-hot encoding)	One-hot encoding will be applied with drop-one to avoid introducing bias into the models and also to avoid multicollinearity

To illustrate the cleaning, here is a before and after distribution of the Parking feature:

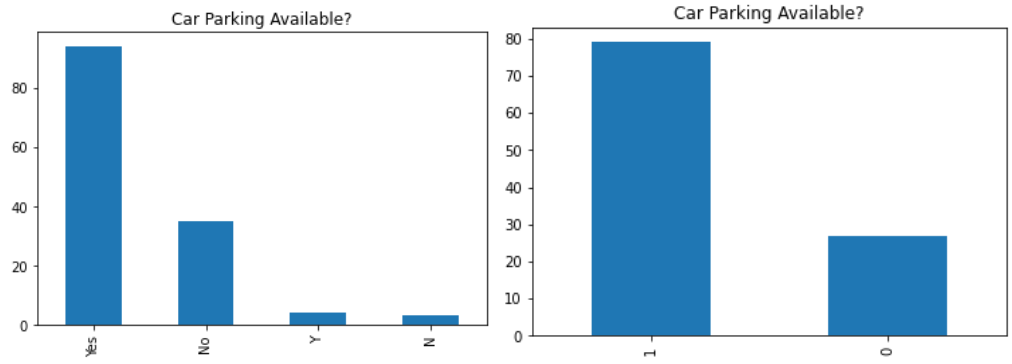

Figure 6: Chart showing before and after cleaning of car parking feature

3.3.FEATURE ENGINEERING

Feature engineering was not required by the client in this example however it is worth mentioning a brief exploration of the possibilities. Whilst the measures of population with a certain driving distance and floor size are absolute it would be interesting to see how these measures interact with each other and the number of staff listed at the store. After all, customer service is a huge part of creating a good impression and increasing the likelihood of a sale. Large populations and large stores that possess few staff may (or may not) suffer from issues such as being unable to replenish stock quickly or not enough staff to service customers. The population to floor space ratio may give a sense of the store being too empty or too crowded. The CRISP-DM model allows for the modeller to circle back to the client to understand if any of these proposed metrics would be of use and possibly even encourage the client to engineer their own features.

3.4.DATA SCALING

The two main types of data scaling are normalisation and standardisation. We know that decision trees are largely unaffected by data scaling whereas logistic regression and MLPs do benefit from feature scaling. To ensure the best results for each model type the baseline model shall be analysed with no scaling, normalisation, and standardisation to see which gives the best baseline model in each case. Multiple datasets will be created to preserve the master dataset.

We would not expect standardisation to provide the optimal baseline as the data does not seem to follow a normal distribution however the experiment shall be performed as described. Similarly, we would not expect the decision tree model to be overly affected by feature scaling as this is not a requirement for this type of model.

3.5.OUTLIER DETECTION

Only two features contain the presence of mild outlying values which are the 10 and 20 minute populations. When looking at extreme outlying values this reduces to the 10 minute population only (3 records). Ideally this would be a conversation with the client however as these values are small and likely to be normal should the client open stores near any the UK's large cities the values will be retained. It is also possible that should these outliers adversely affect the model then feature selection may naturally remove this feature given its high correlation to other similar features.

4. MODELS

4.1.EQUALITY OF OUTCOME

To ensure equality of outcome across the various tests, the random state will be set to 42. Additionally, the same variable containing the cross validation splits will be used for all models to ensure consistency across the model selection process. This is set to use RepeatedStratifiedKFold validation with number of folds equal to 5 and number of repeats equal to 3. The codebase will be setup and tested before results are collected. Once ready, the python kernel will be restarted, and each code cell walked through in sequence to ensure results are collected in sequence from the same test run.

4.2.RANDOM FOREST DECISION TREE

4.2.1. BASELINE MODEL

The random forest model is the preferred decision tree model as it is a versatile and powerful model that makes use of ensemble learning. The model is first tested with no scaling, normalisation, and standardisation of features:

Scaler	Accuracy	F1	Precision	Recall
None	0.648	0.652	0.695	0.678
MinMax	0.657	0.660	0.696	0.673
StandardScaler	0.651	0.667	0.685	0.684

The results show that the normalisation is most effective providing the best accuracy and precision, which was identified as a key metric.

4.2.2. HYPERPARAMETER TUNING

The random forest model was tuned for the following hyperparameters:

criterion	max_features	n_estimators	Accuracy	Precision
gini	sqrt	10	0.625	0.661
gini	sqrt	100	0.657	0.670
gini	sqrt	250	0.667	0.677
gini	**sqrt**	**500**	**0.695**	**0.720**
gini	sqrt	1000	0.663	0.668
gini	log2	10	0.617	0.666
gini	log2	100	0.673	0.695
gini	log2	250	0.661	0.678
gini	log2	500	0.667	0.675
gini	log2	1000	0.676	0.682

criterion	max_features	n_estimators	Accuracy	Precision
entropy	sqrt	10	0.644	0.689
entropy	sqrt	100	0.653	0.668
entropy	sqrt	250	0.676	0.698
entropy	sqrt	500	0.676	0.690
entropy	sqrt	1000	0.685	0.692
entropy	log2	10	0.639	0.673
entropy	log2	100	0.679	0.685
entropy	log2	250	0.685	0.705
entropy	log2	500	0.676	0.697
entropy	log2	1000	0.676	0.684

The results show us that one set of hyperparameters comes out on top for both accuracy and precision and is therefore the clear choice for optimised parameters going forward. If there had not been a clear winner then the resulting trade-off between accuracy and precision would have to be referred to the client as part of the CRISP-DM model framework. Therefore, the best hyperparameters are defined as: criterion: gini, max_features: sqrt, n_estimators: 500

4.2.3. FEATURE SELECTION

The random forest feature importance lists the following 4 features as the most important:

Feature Index	Score	Feature name
Feature 0	0.12625	Staff
Feature 13	0.10674	Clearance space
Feature 1	0.09625	Floor Space
Feature 15	0.08863	Competition score

The model will be trained against with this reduced set of active features to see if any final improvements can be made:

Scaler	Accuracy	F1	Precision	Recall
MinMax	0.635	0.646	0.656	0.618

As both accuracy and precision are worse under the new feature selected model it can only be concluded that the feature selection with tuned hyperparameters did not assist in improving the model.

4.2.4. CONCLUSION

The most promising random forest model is defined as the hyperparameter tuned model with normalisation.

4.3.LOGISTIC REGRESSION CLASSIFIER

4.3.1. BASELINE MODEL

Logistic regression is a commonly used classifier model. The model is first tested with no scaling, normalisation, and standardisation of features. The solver used is liblinear and the penalty is set to L2 regularisation (default C=1):

Scaler	Accuracy	F1	Precision	Recall
None	0.516	0.549	0.519	0.588
MinMax	0.799	0.801	0.821	0.789
StandardScaler	0.773	0.783	0.780	0.790

Normalisation proves to be the best scaling function and yields the highest accuracy and precision with almost identical recall in comparison to standardisation. In this instance, the normalised model is the best baseline model.

4.3.2. HYPERPARAMETER TUNING

The logistic regression model was tuned for the following hyperparameters:

C	penalty	Solver	Accuracy	Precision
100	L2	newton-cg	0.757	0.768
100	L2	lbfgs	0.757	0.768
100	L2	saga	0.764	0.776
10	L2	newton-cg	0.776	0.796
10	L2	lbfgs	0.776	0.796
10	L2	saga	0.776	0.796
1	L2	newton-cg	0.799	0.821
1	L2	lbfgs	0.799	0.821
1	L2	saga	0.799	0.821
0.1	L2	newton-cg	0.774	0.775
0.1	L2	lbfgs	0.774	0.775
0.1	L2	saga	0.774	0.775
0.01	L2	newton-cg	0.692	0.649
0.01	L2	lbfgs	0.692	0.649
0.01	L2	saga	0.692	0.649
100	L1	liblinear	0.757	0.763
10	L1	liblinear	0.764	0.782
1	L1	liblinear	0.764	0.762
0.1	L1	liblinear	0.490	Nan
0.01	L1	liblinear	0.490	Nan

Disappointingly the L1 lasso regularisation failed to make a telling impact with its feature shrinkage. The optimal model appears to be one with **L2 regularisation, C=1 and a choice of solvers** (newton-cg, lbfgs or saga).

4.3.3. FEATURE SELECTION

The logistic regression feature coefficients list the following 5 features as the most important:

Feature Index	Score	Feature name
Feature 0	0.771	Staff
Feature 15	0.475	Competition score
Feature 6	0.421	Location_Shopping Centre
Feature 14	0.371	Competition number
Feature 5	-0.357	Location_High Street

The feature importance coefficients are a crude method of predicting how important is in predicting a zero or 1 outcome. A positive score is associated with predicting 1, whilst negative is associated with predicting a zero.

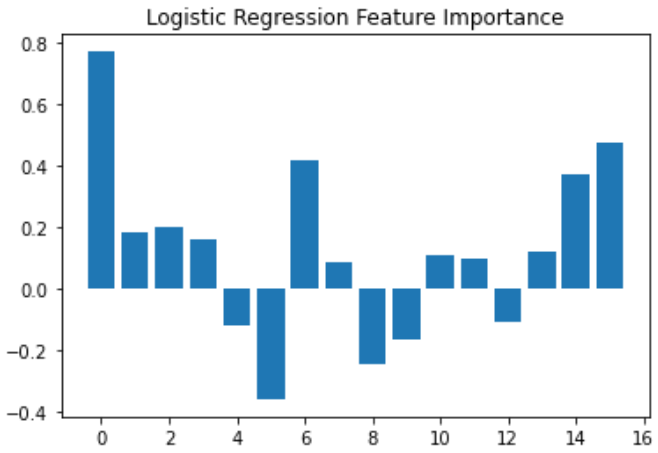

Figure 7: Chart showing importance of features in model

The model will be trained against with this reduced set of active features to see if any final improvements can be made:

Scaler	Accuracy	F1	Precision	Recall
MinMax	0.815	0.822	0.813	0.839

The feature selection appears to have improved the model taking all metrics to over 80%.

4.3.4. CONCLUSION

The most promising logistic regression model is defined as the hyperparameter tuned model with normalisation and feature importance selection enacted (top 5 important features).

4.4.MULTI-LAYER PERCEPTRON (NEURAL NETWORK)

4.4.1. BASELINE MODEL

The Multi-Layer Perceptron model is the preferred neural network choice for this last model selection test. The model is tested with no scaling, normalisation, and standardisation of features with the following baseline parameters:

hidden_layer_size=10, alpha=0.001, batch size=auto, learning_rate=constant, learning_rate_init=0.01, max_iter=1000

Scaler	Accuracy	F1	Precision	Recall
None	0.516	0.549	0.519	0.588
MinMax	0.799	0.801	0.821	0.789
StandardScaler	0.773	0.783	0.780	0.790

The results show that the normalisation once again has the greatest effect on our baseline model, pushing both the accuracy to 79.9% and precision to just over 82%

4.4.2. HYPERPARAMETER TUNING

The random forest model was tuned for the following hyperparameters using the GridSearchCV function with all other hyperparameters being left as per the baseline values:

activation	hidden_layer_sizes	Solver	Accuracy	Precision
relu	25	adam	0.736	0.734
relu	50	adam	0.726	0.724
relu	100	adam	0.755	0.755
relu	25	lbfgs	0.748	0.760
relu	50	lbfgs	0.764	0.763
relu	100	lbfgs	0.745	0.757
tanh	25	adam	0.755	0.756
tanh	50	adam	0.736	0.734
tanh	100	adam	0.745	0.749
tanh	25	lbfgs	0.745	0.748
tanh	50	lbfgs	0.758	0.757
tanh	100	lbfgs	0.732	0.733

On this occasion the hyperparameter search did not yield any improvements to the baseline model. This is most likely to be caused by a poor choice of hyperparameters or an excellent baseline model. A future search would be advised to play with the learning rates, regularisation, batch sizes and other more advanced parameters. The search space analysed here does not show much variation meaning that the parameters explored did not change the overall model in any meaningful way.

4.4.3. FEATURE SELECTION

Feature selection for the MLP will be analysed with a simple wrapper method based on the local search. The mutation operator for this search will be a simple flip operation that takes a random feature and switches it to the opposite of its Boolean value (include -> exclude, or vice versa).

Using our baseline MLP model and the existing cross fold validation variable a baseline precision score of 0.744 was obtained. Over the course of 150 iterations a fit score of 0.849 was observed after 16 iterations and remained static thereafter, using 9 features. The 9 features selected are:

Feature index	Feature name
0	Staff
2	Window
5	Location_High Street
6	Location_Shopping Centre
7	Location_Village
10	20 min population
12	Store age
14	Competition number
15	Competition score

The baseline MLP model was re-run with the 10 selected features to see if any final improvements could be made:

Scaler	Accuracy	F1	Precision	Recall
MinMax	0.697	0.691	0.730	0.666

Feature selection does not appear to have made any improvements to our baseline model, in fact it performs worse than using all features.

4.4.4. CONCLUSION

The most promising neural network model is defined as the baseline model with normalised data scaling.

4.5.FINAL MODEL

The final model will be chosen from amongst the best models from each type:

Model Type	Accuracy	F1	Precision	Recall
Random Forest	0.657	0.660	0.696	0.673
Logistic Regression	0.815	0.822	0.813	0.839
Neural Network	0.799	0.801	0.821	0.789

From the experiments conducted the Logistic Regression model wins on accuracy but the Neural Network wins on precision, by 0.8%. If we look at the F1 score and recall we see that the Logistic Regression model outperforms the Neural Network. Given that the precision difference is < 1% we can say that it is a worthwhile trade-off to select the Logistic Regression model for final testing.

As always with CRISP-DM, if we wanted to be doubly sure that choosing a model with a very slightly lower precision was acceptable, we could circle back to the client and explain our thinking before proceeding with their permission.

The Logistic Regression model (with normalised data, tuned hyperparameters and a subset of features) will be tested against the entire training set without cross validation to see how the metrics hold up:

Model Type	Accuracy	F1	Precision	Recall
Logistic Regression	0.821	0.823	0.807	0.852

The results show that the model performs very closely on the entire training set to the how the model performed against the cross fold validation sets. With the consistency of results, it gives confidence to say that the model is stable and good enough to put forward for testing against the test set data.

A final summary of the model is given as:

Logistic Regression Model: Classifier

Feature Selection (5): Staff, Competition score, Location_Shopping Centre, Competition number, Location_High Street

Hyperparameters: Solver: lbfgs, penalty: L2, C: 1

(Random seed: 42)

5. EVALUATION

To prepare the test dataset, data cleaning and preparation steps were carried out on the test set a were performed on the training set. The data was normalised as per the training set findings (scalar.transform and not .fit_transform).

5.1.TEST SET RESULTS

The results of testing against the **test set** data are as follows:

Model Type	Accuracy	F1	Precision	Recall
Logistic Regression	0.778	0.769	0.769	0.769
Delta vs. training	-0.043	-0.054	-0.038	-0.083

The model performs almost to the standard of the training set model. The accuracy is 4.3% lower and the precision 3.8% lower than in training however it is still reasonably close given the limited data.

5.2.CONFUSION MATRIX

The confusion matrix of 27 test observations is given as:

	Predicted: Bad	Predicted: Good
Actual: Bad	11 (TN)	3 (FP)
Actual: Good	3 (FN)	10 (TP)

TN = True Negative; FN = False Negative; FP = False Positive; TP = True Positive

A summary of confusion matrix statistics shows that whilst the model provides better metrics than simply guessing, there is approximately a 20%, or one in five chance, that a new store will not perform as well as expected (FP) however there is the same chance that a bad store will perform better than expected (FN) so on average the odds could even themselves out. One way to quantitate the risk of false positive would be to investigate whether the cost of an under-performing store is greater than the potential revenue gain of an over-performing store and weigh this against the number of stores that the client is planning to open. False negatives are less of a concern as it will not cost the client money.

Metric	Result	Comments
Accuracy	77.8%	Approximately one out of every five predictions will be correct
Classification Error	22.2%	Approximately one out of every five predictions will be wrong
Sensitivity	76.9%	The model will identify approximately 3 out of every 4 good stores correctly
Specificity	78.6%	The model identifies approximately 4 out of every 5 bad stores correctly
False Positive Rate	21.4%	The model identifies approximately 1 out of every 5 bad stores incorrectly
Precision	76.9%	Approximately 3 out of every 4 stores predicted to be 'Good' will actually be 'Good'

5.3.AUC & ROC

The Area Under Curve statistics is calculated as 77.7% which is in the range of an acceptable score for a model (0.7-0.8 is considered acceptable). The ROC curve, which shows the trade-off between specificity (x-axis) and sensitivity (y-axis) is plotted and is closer to approaching the top-left corner than the no-bias centre diagonal (shown in red) which implies that this is a good model but could be better with the current model performing at the elbow of the blue line:

Figure 8: Chart showing Receiver Operating Characteristic (ROC) curve of final model

5.4.MODEL ERRORS

The model made 6 errors from the 27 test observations. Three were False Positive and three were False Negative. After extracting the errors, it may be possible to see a pattern in the data that is causing the issue with the selected features (Location data has been summarised for space saving):

Index	Staff	Location	Comp Score	Comp Number	Actual Perform	Predicted Perform
15	6	Shopping Centre	12	11	1	0
18	5	(Not selected)	15	15	1	0
21	8	High Street	12	14	1	0
24	8	High Street	14	17	0	1
25	6	High Street	15	17	0	1
26	5	High Street	13	17	0	1

It is hard to say with certainty, but all the stores that were incorrectly predicted to be 'Good' are located on the high street with a competition number towards the upper end of the scale for that feature. It is beyond the scope of this assignment however perhaps better results could be gained by analysing the threshold probability for predicting 'Good' or 'Bad' and optimising it to reduce these sorts of errors instead of diving straight into remodelling. The threshold for binary classification defaults is usually 0.5 (above = 1, below = 0) so there is likely scope for analysis and improvement here.

5.5.DEPLOYMENT & SUMMARY OF CRISP-DM

At each stage of the CRISP-DM process we have seen issues occur that would require input or clarification from the client in a real-world scenario. This highlights the importance of good communication and a good working relationship with any client that engages you to perform a data driven project (or any project for that matter).

Now that the model has been produced it would be delivered to the client ideally with some form of presentation or walk-through so that they can understand the model clearly and have a good grasp of the strengths, weaknesses, and limitations particularly in relation to the tricky issue of False Positives and minimising the potential for the client to invest substantial sums of money into a store that may fail.

We have seen the model perform acceptably on a limited dataset. It would be a good idea to ask the client for a second tranche of data so that the model can be retrained and improved upon against a larger dataset. The one provided only contained stores in towns beginning with the letter 'S' so clearly the client has more data available.

Should the model be accepted by the client and deployed an aftercare package will need to be agreed to support or handover the model to the client. The nature of data driven models is that as data changes over time, it will almost certainly be the case that the model will need to be updated or tuned as the data develops. A big risk is changes to the underlying data structures, data values or the rate of data consumption. Degradations in performance can be proactively managed through development of appropriate Business Intelligence solutions as well as 'monitor and alarm' processes.