Assignment ITNPBD6
Student ID 2727809

Abstract

We are living in an era of datafication, which means that almost everything around us is becoming a data source. People try to receive data from every possible resource because data itself is a new level of modern intelligence and our task is to get this intelligence from provided data. In order to get that insight from the data, we have to implement a broad range of techniques and algorithms.

Generally speaking, data mining is the process of learning the current world to predict the future. The main idea of data mining assumes that patterns which could be discovered in the existing data will exist in the future as well. As without data mining, it is absolutely impossible to understand which data (variables) could be useful for getting valuable insight, people have to collect a lot of redundant data before data mining techniques will be applied.

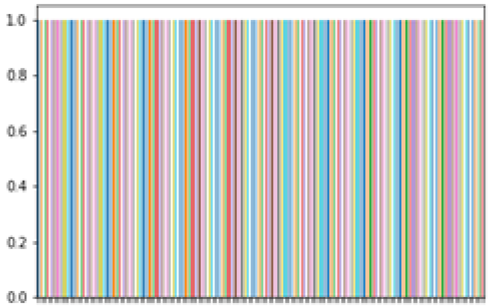
Application of data mining in different fields can bring us a lot of innovations and improvements. One example is illnesses prediction and its prevention in medicine, fraud detection in banking, criminal recognition in security and so on. When it comes to the commercial sector, entrepreneurs are collecting data about their current businesses in order to increase the profit of the company in the future. By understanding which factors influence profit, they can manage their current business processes and implement the knowledge for future business projects. For instance, data mining can discover a relationship between the location of the store and its profit, which is extremely important for entrepreneurs when they run new business projects.

Our client is the owner of over 100 shops throughout the UK. Histeam managed to collect some data to analyze.The data set consist of 20 variables (columns) and 136 data points (rows). Our goal is to understand through applying different machine learning algorithms and techniques which variables most influence profit and performance of the shops. That allows a business owner to focus on a specific part of hisbusiness and, by changing them, increase profit and performance. Moreover, as data collection is a very resource demanding process, by understanding which variables are useless, we additionally save money for the company eliminating them from the future data collection process.

Description of the process

As we mentioned in the previous chapter, our data set contain 20 variables, where 18 of them are independent and two are our goal variables. Before we begin the data mining process, we have to understand which variable we are going to use. The variable selection process is the preliminary phase of each data mining project for two main reasons. Using too few variables leads to missing valuable information. Whereas, using too many variables leads to a very complex model and more likely an overfitted one. Also, due to our data set being relatively small, by using too many variables we have a risk of getting a poor generalized model.

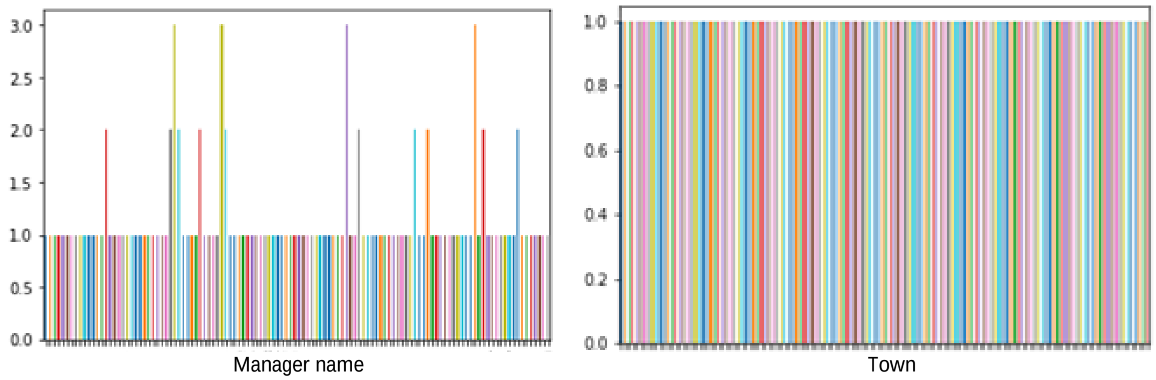
There are several approaches to reduce the number of variables, also known as “Dimensionality reduction”, such as Filters, Wrappers, and Embedded.Start off by Filters. One of the most well-known ways to look at the data is visualization. By doing so we are able to understand how our data is distributed and identify outliers, minority values, flat and wide variables, and data entry errors.The first variable, which is worth considering is “Store ID” in picture 1.1:



1.1 Distribution of “Store ID” variable

Looking at the picture, we can see a vivid example of flat, wide distribution, where each value is unique. This variable is useless for data mining because it is impossible to find a pattern from such data. As a result, we will remove the “Store ID” variable from the data set.

The similar situation with a distribution of variables “Town” and “Manager name” in picture 1.2.



1.2 Distribution of "Manager name" and "Town" variables

Some peaks in the "Manager name" bar chart means that the names of some managers are the same, which leads to counting that values. In real life, in such a vague situation we should clarify the information, but we assume that all managers are different persons. As a result, both variables have flat and wide distribution with unique values and should be deleted from the model data set.

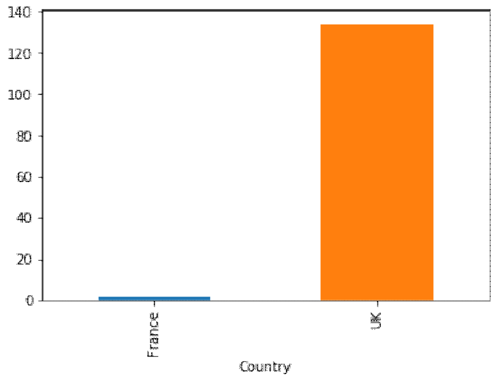
One more reason to delete numerical variable is its very low correlation coefficient with the dependent variable, in our case is profit. Table 1 below shows us correlation coefficients for each variable.

Profit	1.000000
Competition score	0.470811
Window	0.323464
Clearance space	0.319751
Floor Space	0.313713
Competition number	0.212082
Staff	0.147357
40min population	0.017691
20 min population	0.012792
30 min population	-0.012406
10 min population	-0.030268
Store age	-0.099801
Demographic score	-0.134655

Table 1: correlation coefficients of numerical variables

As we can note inTable 1, variables “20 min population” and “30 min population” have extremely low coefficients. So, they might be eliminated. However, during data mining, we are going to use additional feature selection methods such us “Wrappers” and, as a result, we let them participate in the data modeling in order to double check our assumption.

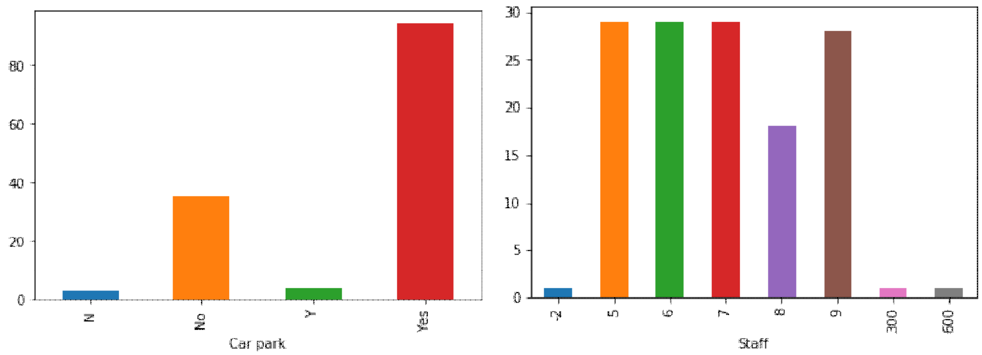
The last variable, which deserves our consideration before the data mining process is started, is “Country”. As there are only two values from 136 overall are France and the rest are the UK. Moreover, both correspondents for the France towns actually locate in the UK, which gives us the tip that values “France” was entered wrongly. So, we eliminate this variable for the same reason as for the “Store ID”, “Town” and “Manager name”.



1.3 Distribution of country variable

Data pre-processing

By observing the rest data, we have noticed that the variable “Car Park” contain some entry errors. To be more precise, the value “No” was mistakenly entered as “N” and value “Y” instead of “Yes”. It is obvious entry errors, which was solved by replacing incorrect values by correct in Python.



1.4 Distribution of "Car park" and "Staff" variables

The variable "Staff" also contains some problems. One value was negative, that impossible for a number of employees. Also, two values were extremely different from the others. It could be considered an entry error as well as for "Car park". We already mentioned before that for such cases, the best way to solve that problem is to clarify this information with business, but as we do not have that opportunity, we should eliminate these values from the data set.

Also, the variable "Location" contains only one value "Village". Despite that, it could be data entry or outlier it is definitely "minority value" that could not contribute a lot to the model. As a result, we removed that value from the data set as well. Table 2 below depicts the summary of the variable selection process and the final composition of them for the data mining process.

Variables	Candidate for a model	Type	Manipulation
Town	No	Nominal	Removed variable
Country	No	Nominal	Removed variable
Store ID	No	Numeric	Removed variable
Manager name	No	Nominal	Removed variable
Staff	Yes	Numeric	Remove negative value and outliers
Floor Space	Yes	Numeric	
Window	Yes	Numeric	
Car park	Yes	Nominal	Correct entry errors
Demographic score	Yes	Numeric	
Location	Yes	Nominal	Remove value "Village"
10 min population	Yes	Numeric	
20 min population	Yes	Numeric	
30 min population	Yes	Numeric	

40min population	Yes	Numeric	
Store age	Yes	Numeric	
Clearance space	Yes	Numeric	
Competition number	Yes	Numeric	
Competition score	Yes	Numeric	

Table 2 Summary variable selection and preparation table

One more very important thing we should not forget to do is nominal variables encoding. This is due to the fact that especially regressions models (decision tree we will not need such encoding) require variables to be numeric. In our case, we have variables such as “Location” and “Car park” which are nominal, so that need to be encoded into dummy variables. We cannot just assign numeric values for a particular category like {Retail Park = 1, Shopping Centre = 2, High Street = 3} because the model will measure a numerical distance between values, which have no sense.

Techniques

There are several techniques we are going to implement to predict profit and performance.For the profit, it is Multilinear regression and Multi-layer Perceptron.For the performance, it is Logistic regression, Multi-layer Perceptron, and Decision tree.

Multiple Linear Regression:

Multiple Linear Regression assumes linear dependency between two or more independent variable and a predicted (dependent) variable by fitting a linear equation to data. In other words, multiple linear regression attempts to find dependency and association between values of predictors (independent variables) $X1$, $X2$, … and value of predicted variable Y.The general equation for the multiple linear regression is:

$$Y = \beta0 + \beta1X1 + \beta2X2 + \cdots \beta kXk + e, \text{ (Equation 1)}$$

Where Y – is our predicted variable, $X1$, $X2$, …Xk - are predictors,$\beta0$ – is an intercept,$\beta1$,$\beta2$ – are slope or gradient. In our case, Y is profit, this is the parameter that we are going to predict as a result of our model. The X1, X2, …Xk are our independent variables from Table 2 (Floor Space, Car park and so on). The β1,β2, …βk – can also be considered as an independent contribution of each variable.

There are several learning algorithms for multiple linear regression such as the Stochastic Gradient Descent (SGD)and the most popular is the Ordinary Least Squares (OLS). The OLS defines its cost function finding the least squares estimator for β. One more criterion for successful multiple linear regression learning is the right set of variables. The Stochastic Gradient Descent also can be used to estimate the parameters for linear regression models. It uses an iterative approach which decreases the error function by making one step at a time. Each step is guided by getting of the error function and limited by learning rate. The smaller steps have a higher chance of notmissing the smallesterror butrequires a longer training time.

For all models and for multiple linear regression particularly it is very important to find the best variable set. Even after applying “Filters” method as an initial part of variable selection process, we have a set of variables which we consider as an initial set before an additional variable selection method will be applied during model learning. While finding the best model we willadd or remove variables from the initial set to find a set of variables which gives us the best accuracy of the model. As part of the process of fitting the model, we can implement scaling the data process, because our variables have different ranges. One more reason to scale the data is that variables with very small values would have a small impact on the target variable, so we need to try it while modeling.

When it comes to a result representation, multiple linear regression is easy to understand. As we mentioned before, Equation 1 represents a set of parameters, which shows us how much each of them contributes to the output. The sign of parameter gives us the direction of the changes that would influence the result. Also, the care is needed if the input data were scaled because in this case, the values would be in not original ranges.

If we speak about multiple linear regression, the most common measure of the result is the correlation coefficient, which presents the strength of the relationship between the set of independent and dependent variables. The value of the correlation coefficient is between -1.0 and 1.0. The negative sign shows the negative relationship and 0.0 shows no relationship between variables.

We also have to take into account the error. We are going to consider the root mean square error (RMSE). The RMSE shows us the standard deviation of the prediction errors and measures how far from the regression line data points are.

Multilayer Perceptron:

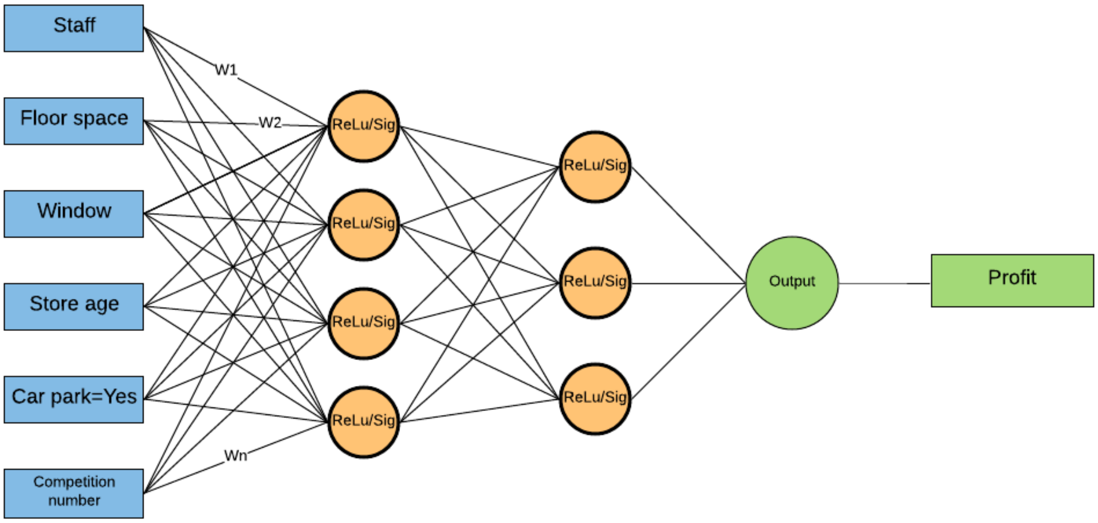
Multilayer Perceptron is an artificial neural network, which is usually used for supervised machine learning to learn dependency between inputs and output. It consists of an input layer to receive the data, an output layer that produces the prediction about the input layer, and an arbitrary number of hidden layers between them.Every layer is fully connected to the next layer, where every connection is weighted, and hidden layers and output once have bias neurons.
The model learns by using the backpropagation algorithm. For each training instance, the algorithm feeds it to the network (with random weight) and calculate the result of every neuron in each layer (forward pass) by passing through the activation function. The output layer forms the output for the function (single value). Then, by measuring the output error, it can calculate how much each hidden layer contribute to the output error. For this reason, the algorithm makes forward pass until it reaches the input layer. Finally, the backpropagation algorithm (see Equation 2) slightly changes weight to reduce error (Stochastic Gradient Descent step).Each node can have different activation function, but in reality, inputs have linear function and hidden layers have Rectified linear (ReLU) and sigmoid.

$$W_{next} = W + {}_{delta}W$$

$${}_{delta}W = \text{-learning rate*gradient + momentum*}{}_{delta}W \text{ previous. (Equation 2)}$$

For multilayer perceptron, the most common hyperparameters arethe number of hidden layers, momentum and learning rate. Care is needed in this step because a great number of hidden layers leads to higher variance and lower bias, which means that we risk getting overfitted model. Low bias means low training error, whereas high variance means high test error. This is why cross-validation is very important for multilayer perceptron modelling.At a data preparation phase, for multilayer perceptron we have to transform nominal variables to dummy variables. The opposite site of such recoded is increasing number of input nodes and its correspondent weights. As a result, we have to be especially careful controlling bias and variance errors.

The picture 1.5 below depict the example of multilayer perceptron for profit prediction. It consists of six input variables and two hidden layers (one of 4 neurons, another of 3).



1.5Multilayer perceptron model for profit prediction

Logistic Regression:

Logistic Regression is commonly used for probability prediction that an instance belongs to a particular class. If the probability greater than 50% means that the instance belongs to the class, if it is less than 50%, then does not belong (binary classification). The main thing that we have to take into account that we cannot present result a probability.To be more precise, logistic regression is useful method when the output has nominal values. Like the Linear regression, logistic regression calculates the weighted sum of input variables plus error, but it does not provide result directly, is emit the result in form of logistic. Logistic is a sigmoid function, output of which is number from 0 to 1. Once the logistic regression provides the probability it is easy to decide if the output belong to positive class or not. It is 0 if $p < 0.5$ and 1 if $p > 0.5$.

Decision tree:

Decision tree is a machine learning algorithm that can perform either regression or classification tasks. It works better with nominal data, but numerical data is also possible to process by splitting into bins. As the decision tree is very intuitive and its result easy to understand it is called “White box”. In contrast with MLP, which is often called “Black box” for its complexity.As a tree in the real life, the decision tree has a tree-like decision model.

The root of the tree is its top node. A good example of algorithm for choosing the top node is ID3. Its goal is to remove the most uncertainty about the class and divide the data into smaller data sets. Each internal node contains one variable, base on which the tree split into branches (edges) and a class label on a leaf node. Then each of the variable from the subset perform the same action as root node above, remove uncertainty and divide data into smaller sub sets. The process repeats, and the tree grows until all instances of the branch are in the same class.

While choosing the new node, the algorithm evaluates all the variables and pick the best one, which minimize the cost function and give the greatest information gain. The main measure of information is its probability that particular event occurs. The sum of the probabilities of each event multiplied by its value is entropy, which is the main measure of uncertainty. Moreover, as entropy is measure of uncertainty of particular variable, the more uniform distributed the variable, the greater entropy gets.

The process to achieve the solution

The most commonprocess planer for data mining processes is the Cross Industry Standard Process for Data Mining (CRISP-DM). It includes such phrases as business understanding, data understanding, data preparation, modeling, evaluation and, as a result of our work – deployment.

The initial part is to understand the task we have to perform and which variable(s) we are aimed to consider as a target. As we said before, our goal is to predict profit and performance from the given data. In our case we have a written description of the task, but in real life we also have an opportunity to talk with a customer in order to clarify all necessary information.

If during the task understanding we have no issues, we have some questions arising during the data understanding phase about data quality. Without an opportunity to contact the business to ask any questions, it was done as part of data pre-processing phase.Before the phase of data modeling was started, we randomly split all the data into two parts. One is a training set, and another is a test one.

The training set is 80% of the data, which is 106 rows and test set is the rest 26 rows. The training set we will use to train model, whereas the test set we will use to test our model and it will be left aside until the final stage. As an intermediate model evaluation process, we will also use validation set, which is 20% of training set.As well as data quality is not perfect, data quantity also poor, just 132 data points in total after data pre-processing. Also, as we have some nominal variables, data encoding was implemented by creating dummy variables.

As we stated before, for data modeling we will use ensemble method consist of the following machine learning algorithms:

Profit:

Multiple linear regression

Multilayer perceptron

Performance (classification):

Logistic regression and

Multilayer perceptron

Decision tree

As we stated in the data Chapter 2.2 we removed some variables from the initial data as a dimensionality reduction measure. The rest of the variables participated as an input variable for model fitting. Initially we run all models with the default parameters. After getting the results, we enter them into a corresponding result table. There are two tables for both target variables.We continued working with the same model by changing hyperparameters and feature selection in order to get better results.In Tables 3 and 4 are the results getting by using test set are in bold fond.

Algorithm	Accuracy (Correlation coefficient)	Root mean square error	Dummy rescale	Feature selection	Data split	Variables	Hyperparameters
Multiple Linear regression	0.862	401978	Yes	No	80%	All	
Multiple Linear regression	0.8829	360235	Yes	Yes/ Eliminate colinear attribute /Greedy	80%	Staff, Window, Car park=Yes, Location=Shopping Centre, Competition number, Competition score	
Multiple Linear regression	0.7799	437335	Yes	Yes/ Eliminate colinear attribute /Greedy	Cross Validation	Staff, Window, Car park=Yes, Location=Shopping Centre, Competition number, Competition score	
Multiple Linear regression	**0.582**	**619838**	Yes	Yes/ Eliminate colinear attribute /Greedy	20% Test set	**Staff, Window, Car park=Yes, Location=Shopping Centre, Competition number, Competition score**	
Multi Layer Perceptron	0.4308	811062	Yes	No	80%	All	
Multi Layer Perceptron	0.8218	416793	Yes	Yes	80%	Staff, Window, Car park=Yes, Location=Shopping Centre, Competition number, Competition score	
Multi Layer Perceptron	0.8852	333173	Yes	Yes	80%	Staff, Window, Car park=Yes, Location=Shopping Centre, Competition number, Competition score	1 Layer-6 neurones/ Momentum =0.4/ Learning rate =0.005
Multi Layer Perceptron	0.7906	426754	Yes	Yes	Cross Validation	Staff, Window, Car park=Yes, Location=Shopping Centre, Competition number, Competition score	1 Layer-6 neurons/ Momentum =0.4/ learning rate =0.005
Multi Layer Perceptron	**0.5793**	**619014**	Yes	Yes	20% Test set	**Staff, Window, Car park=Yes, Location=Shopping Centre, Competition number, Competition score**	**1 Layer-6 neurones/ momentum=0.4/ lerning rate =0.005**

Table 3. Resultsof regression modeling

Algorithm	Accuracy (correctly class instances)	Root mean square error	Dummy rescale	Feature selection	Split	Variables	Hyperparameters
Logistic Regression	61.9048 %	0.3879	No	No	80%	All	Default
Logistic Regression	61.9048 %	0.3879	No	Yes	80%	Staff, Window, Car park, Location, Competition number, Competition score	
Logistic Regression	**53.8462 %**	**0.3881**	No	Yes	20% Test set	Staff, Window, Car park, Location, Competition number, Competition score	
Multi Layer Perceptron	61.9048%	0.4036	No	Yes	80%	Staff, Car park=No, Location=Retail Park, Location=Shopping Centre,Location=High Street, 40min population, Clearance space, Competition number, Competition score	1 Layer-2 neurones/ Momentum =0.3/ Learning rate =0.1
Multi Layer Perceptron	48.1132%	0.4378	No	Yes	Cross Validation	Staff, Car park=No, Location=Retail Park, Location=Shopping Centre, Location=High Street, 40minpopulation, Clearance space, Competition number, Competition score	1 Layer-2 neurones/ momentum =0.3/ learning rate =0.2
Multi Layer Perceptron	47.619%	0.4005	No	Yes	80%	Staff, Window, Car park=No, Demographic score, Location=Retail Park, Location=Shopping Centre, Location=High Street, 10 min population, Store age, Competition number, Competition score	1Layer-2 neurones/ Momentum =0.1/ Lerning rate =0.1
Multi Layer Perceptron	**50%**	**0.451**	No	Yes	20% Test set	Staff, Window, Car park=No,Demographic score, Location=Retail Park,Location=Shopping Centre, Location=High Street, 10 min population, Store age, Competition number, Competition score	1Layer-2 neurones/ Momentum =0.1/ Lerningrate =0.2
Decision Tree	47.619%	0.4892	No	Yes	80%	Staff, Car park, Demographic score, Location, 40min population, 30 min population, 20 min population, Store age, Clearance space, Competition number, Competition score	

Decision Tree	**46.1538%**	**0.4627**	No	Yes	20% Test set	Staff, Car park, Demographic score, Location, 40min population, 30 min population, 20 min population, Store age, Clearance space, Competition number, Competition score	

Table 4. Results of classification modeling

Tables 3 and 4 consist of results and hyperparameters for each regression and classification methods.The Pearson Correlation coefficient (R) was used as an accuracy measure for regression models. By calculating the coefficient of determination (R^2) we can describe the proportion of variance for a dependent variable by independent once. Also,the root mean square error (RMSE)was measure of model error. The RMSE depict the difference between values predicted by a model and observed values. So, the higher the Pearson correlation and lower the RMSE, the better model. Also, for a multiple linear regression model we have the equation to calculate prediction for the profit.

When it comes to classification accuracy measure, we use confusion matrix, which is a table describing performance classification model. The main idea is to test classes on the test data finding True positive, True negative, False positive and False negative results. The final measure calculates the proportion of Correctly Classified Instances, measured in percentages.In Table 5, it depicts a confusion matrix for classification problem, using the test set.

```
a b c d   <-- classified as
1 1 1 0 | a = Good
0 6 0 0 | b = Excellent
2 0 1 3 | c = Poor
0 1 3 2 | d = Reasonable
```

Table 5. Confusion matrix for classification test set.

As we said earlier, we split our data into proportion of 20/80 to be able to test the best model after fitting process. The main point for that is to measure how well our model can be adapt to the new, unseen data. Looking at the results, we can note that accuracy and errors of all models using the training data set are much higher than accuracy and errors the same model using test data. The reason for that is that by training model and tuning parameters we are usually overfit it by getting low bias and high variance. In such case we have to find bias/variance trade-off. The final result, which we get using the test data set, describes how our model generalized and how close the result which we get applying future data. As a result, the best model is the model with the high accuracy and the lowest error.

Analysis

Table 6contains the results of the best models:

Model	Algorithm	Accuracy	RMSE	Variables	Hyperparameters
Regression	Multi Layer Perceptron	0.5793	619014	Staff,Window, Car park=Yes, Location=Shopping Centre,Competitionnumber,Competition score	1 Layer-6 neurones/ momentum=0.4/ lerning rate = 0.005
Classification	Logistic Regression	53.8462 %	0.3881	Staff, Window, Car park, Location, Competition number, Competition score	

Table 6. Final models' results

For regression model, the winner is Multilayer Perceptron with 0.5793 accuracy. The figure 0.5793 is a correlation coefficient which measure statistical correlation between variables in the column “Variables” and “Profit”. RMSE is the most common error measure and it is square root from mean square error to get the same dimension as a predicted value. It depicts how far from the regression line data points are. In our case the error is 619014 which is lower than an allowed threshold of one million. The final variable set is depicted in the “Variables” column shows that we managed to significantly decrease number of variables in comparison with initial variable set. When it comes to hyperparameters, MLP provide the widest range of them, which influence significantly on the result. For our data the most effective hyperparameters are one hidden layer with six neurons with momentum of 0.4 and very small learning rate of 0.005. As we mentioned before, we have gotten much higher results on the training data set but it was an overfitted model with lower generalization. To make the model more generalized, we needed to find trade-off between the results getting from training and test sets (between bias and variance).

The classification model provides lower accuracy but the errors are also smaller. The accuracy is 53.8462 %. This figure relies on a confusion matrix. As we mentioned in the previous chapter, the confusion matrix counts the number of correct and incorrect classes in the test set. In our case, the test set consist of 26 instances, which is relatively low, 14 of them where predicted correctly and 12 incorrectly. The RMSE is 0.3881, which is the lowers error in comparison with other models.

Recommendations for variables

As we can see in Table 6, the number of variables participating and giving the best models result is the same both models. Moreover, this number is much smaller than initial one. As we described in the data preparation chapter, we have eliminated the variables “Town”, “Country”, “Store ID” and “Manager name” at the beginning for their flat and wide distribution. Later on, as a feature selection process while model fitting, we implement “Greedy” and “Eliminate colinear attribute” methods for regression and Backward for classification models. As a result, we found that variables “10,20,30,40 min population”, “Demographic score”, “Floor Space”, “Store age” and “Clearance space” do not positively influence the result and vice versa cold be considered as a noise.

As we can see, the sufficient for a successful model are “Staff”, “Window”,“Car park”,“Location”, “Competition number” and “Competition score”. However, as a general recommendation for the future data collecting, I would keep collecting the data for these variables as well as looking for new ones, which allows me to get higher results.

Recommendations for techniques

The winners in our project are the MLP and Logistic regression for regression and classification model respectively. Both of them provided the highest accuracy and the lowest errors. However, for regression model the Multiple Linear Regression algorithm gave the result which is slightly worse than MLP. Due to our data set consisted of just 132 instances, which is considered as a very small sample, it worth keeping in mind that Multiple Linear Regression is also good candidate for regression prediction.The equation for profit prediction is as follows:

Profit = -3668137.8279 + 176713.4496 * Staff + 23970.7578 * Window + 524794.4999 * Car park=Yes + 431541.426 * Location=Shopping Centre + 42837.9438 * Competition number + 107084.4561 * Competition score

As a final recommendation, the best options for regression are MLP and Multiple Linear Regression and for classification is Logistic Regression. To predict Profit and Performance with higher accuracy it would be better to find new variables that can bring more insight to the data mining.