|  |
| --- |
| Big Data with H2O |
|  |

|  |  |
| --- | --- |
| 10.15.2018 | Assignment 2 |

|  |
| --- |
| For this assignment, I have used 3 datasets for regression, Classification and Multiclass Classification.  They are as listed below:   1. Regression: [Big Mart sales prediction](https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/) 2. Classification: [Genetic Variant Classification](https://www.kaggle.com/kevinarvai/clinvar-conflicting) 3. Multiclass-Classification: [Car Evaluation Dataset](http://archive.ics.uci.edu/ml/datasets/Car+Evaluation) |

Big Data with H2O

Assignment 2

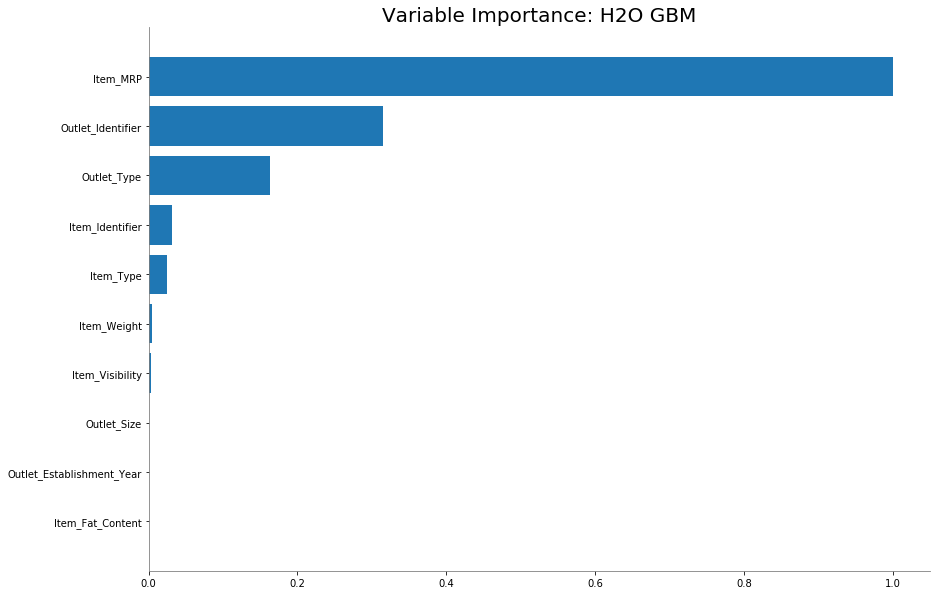
# Regression:

In the Bigmart sales prediction, we are prediction the Item\_Outlet\_Sales for the store. Using H2O AutoML, the best results were observed in Gradient Boosting model. The tabulated results for the same is as shown below:

|  |  |
| --- | --- |
| model\_id | mean\_residual\_deviance |
| GBM\_grid\_0\_AutoML\_20181015\_162406\_model\_12 | 1179801.926 |
| StackedEnsemble\_BestOfFamily\_0\_AutoML\_20181015\_162406 | 1191154.045 |
| DeepLearning\_0\_AutoML\_20181015\_162406 | 1248706.73 |
| XRT\_0\_AutoML\_20181015\_162406 | 1268780.391 |
| DRF\_0\_AutoML\_20181015\_162406 | 1309372.604 |
| GLM\_grid\_0\_AutoML\_20181015\_162406\_model\_0 | 2935313.309 |

## Variable Importance Plot:

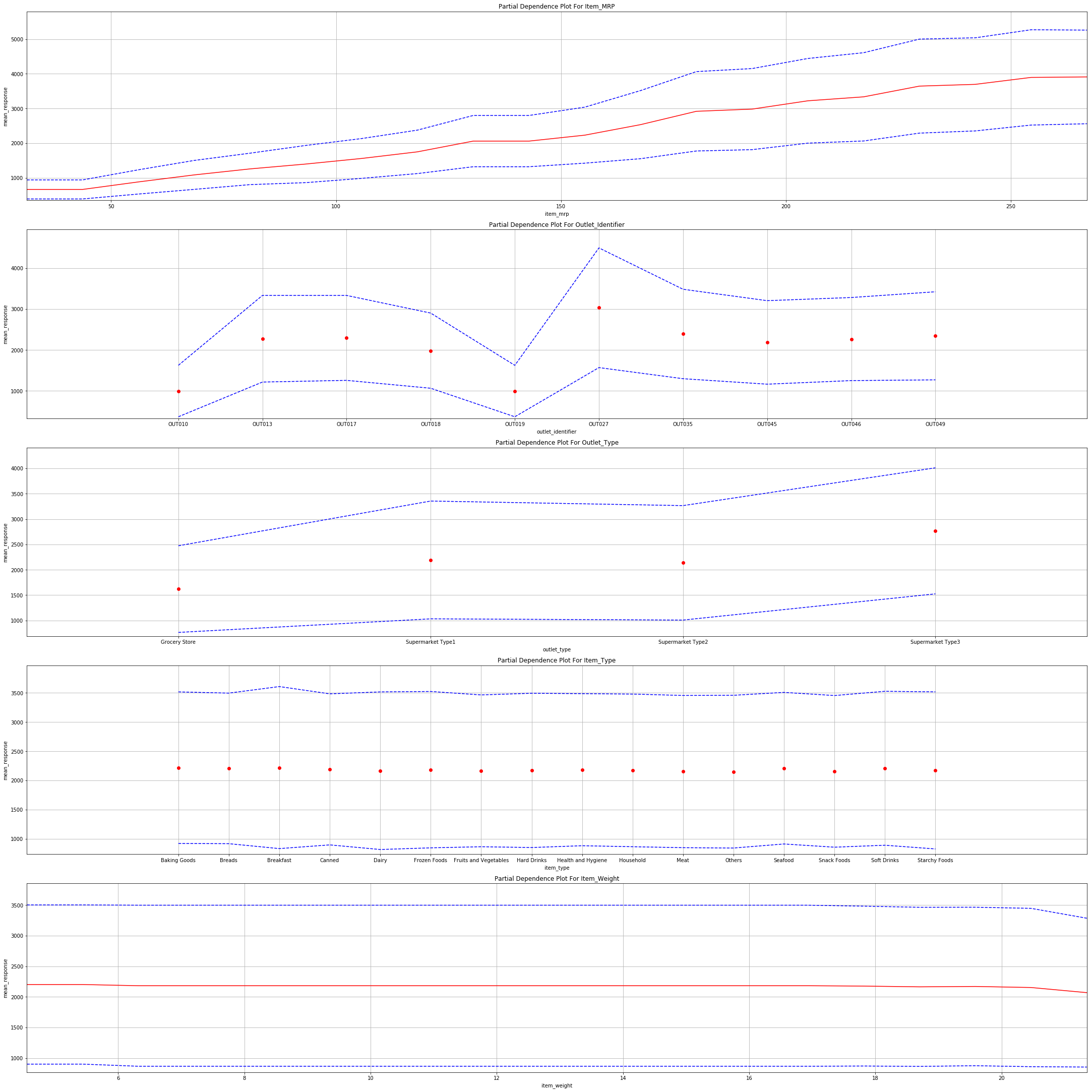
The variable importance lot for the dataset is as shown below:



As we can see, Item\_MRP is the variable which the target variable is most dependent on. On the contrary, variables like Outlet\_Size, Outlet\_Establishment\_Year and Item\_Fat\_Content are not contributing to the model prediction. Hence, these variable can be removed for future processing of data

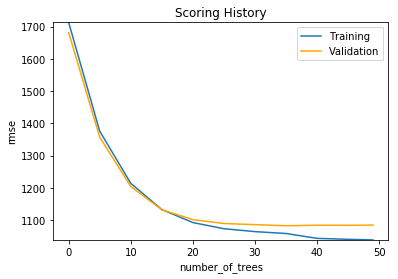
Partial Dependence plot:

The partial Dependence plot for the Big mart sales is as shown below:



As we can observe from the plot, Item\_Outlet\_Sales has an effect from Item\_MRP, Outlet\_Identifier and Outlet\_type. There is very less variation in the plot with respect to Item\_type and Item\_Weight

The scoring history plot for regression is as shown below:



The metrics that were observed were as follows:

|  |  |
| --- | --- |
| MSE | 1175973.309804499 |
| MAE | 758.2964433108918 |
| Mean Residual Deviance | 1175973.309804499 |
| RMSE | 1084.4230308345996 |
| RMSLE | 0.5611744952343632 |
| R2 | 0.6324520757112159 |

## Prediction:

The prediction for the test dataset was performed. The results are tabulated in the csv file by the name “r9ze6Hr2T\_predictions.csv”. The MAE observed for the prediction is: 763.6011906564054.

# Classification

Just as in Assignment 1, For classification, we will continue with the ClinVar dataset to predict the if the variants are likely to have conflicting classifications. For this dataset, The model performance was as follows:

|  |  |
| --- | --- |
| model\_id | auc |
| StackedEnsemble\_BestOfFamily\_0\_AutoML\_20181015\_002205 | 0.822249 |
| XRT\_0\_AutoML\_20181015\_002205 | 0.812527 |
| DRF\_0\_AutoML\_20181015\_002205 | 0.807932 |
| GBM\_grid\_0\_AutoML\_20181015\_002205\_model\_2 | 0.796606 |
| DeepLearning\_0\_AutoML\_20181015\_002205 | 0.766022 |
| GLM\_grid\_0\_AutoML\_20181015\_002205\_model\_0 | 0.762812 |

We will be choosing Distributed Random forest to perform our validations.

## Variable Importance Plot:

## C:\Users\raksh\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\5C326649.tmp

As observed in the graph, most of the variables are contributing to prediction of the CLASS. Amino\_acids are the most important feature for this data. BAM\_EDIT is a feature which does not have any effect on the prediction of the CLASS.

## Partial Dependence Plot:

The partial dependence plot for the classification dataset is as shown below.

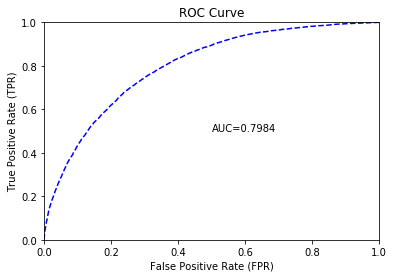
As most of the features in this dataset had more that 20 Unique values, It is not possible to plot the partial dependence plot in H2O. These features are not showing effect on the prediction as they are of low importance and are not effecting the target variable as much as the others.

## C:\Users\raksh\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\E919283F.tmp

The confusion matrix for the classification is as shown below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | Error | Rate |  |
| 0 | 18395 | 5873 | 0.242 | (5873.0/24268.0) | |
| 1 | 2501 | 5377 | 0.3175 | (2501.0/7878.0) | |
| Total | 20896 | 11250 | 0.2605 | (8374.0/32146.0) | |

The ROC curve for the classification is as shown below:

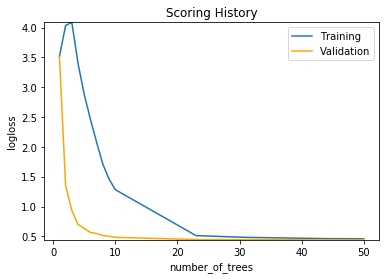


As we can see, the model is giving an AUC of 0.79. The max precision observed was 0.99 and the max recall observed was 0.00001.

These are the metrics that were observed:

|  |  |
| --- | --- |
| MSE | 0.144927314659928 |
| RMSE | 0.38069320280237207 |
| LogLoss | 0.4857284896768785 |
| Mean Per-Class Error | 0.27695568541905735 |
| AUC | 0.7983801242393007 |
| Gini | 0.5967602484786014 |

The scoring history with respect to the logloss is as shown below:



## Prediction:

The prediction for the classification is present in the file 6S2O5A9iv\_predictions.csv.

The precision for the test data was observed to be 0.89.

# Multiclass Classification

For multiclass classification, we are predicting if the car that is being sold is unacceptable/acceptable/good/very good. Using AutoML, we observed that Deep Learning would be the appropriate tool to predict the class of this dataset. The classification leaderboard is as shown below:

|  |  |
| --- | --- |
| model\_id | mean\_per\_class\_error |
| StackedEnsemble\_BestOfFamily\_0\_AutoML\_20181015\_182139 | 0.084032 |
| GBM\_grid\_0\_AutoML\_20181015\_182139\_model\_4 | 0.090978 |
| DRF\_0\_AutoML\_20181015\_182139 | 0.166576 |
| GLM\_grid\_0\_AutoML\_20181015\_182139\_model\_0 | 0.239045 |
| XRT\_0\_AutoML\_20181015\_182139 | 0.247325 |
| DeepLearning\_0\_AutoML\_20181015\_182139 | 0.346026 |

## Variable Importance Plot:

The variable importance plot for the model is as shown below. As we can see, all the features are having a very high effect on the prediction of class.

## C:\Users\raksh\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\204D8C1.tmp

## Scoring History and metrics:

## C:\Users\raksh\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\D469D977.tmp

The scoring history plot is as shown above. The misclassification error was observed to be 0.065.

The metrics are as shown below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| acc | good | unacc | vgood | Error | Rate |
| 273.0 | 2.0 | 13.0 | 0.0 | 0.0520833 | 15 / 288 |
| 11.0 | 34.0 | 0.0 | 7.0 | 0.3461538 | 18 / 52 |
| 34.0 | 0.0 | 820.0 | 1.0 | 0.0409357 | 35 / 855 |
| 3.0 | 10.0 | 0.0 | 37.0 | 0.26 | 13 / 50 |
| 321.0 | 46.0 | 833.0 | 45.0 | 0.0650602 | 81 / 1,24 |

|  |  |
| --- | --- |
| MSE: | 0.05596219173676882: |
| LogLoss | 0.18207870584127683 |
| RMSE | 0.23656329329963433 |
| Mean Per-Class Error: | 0.17479321300044984 |

The predictions for multiclass Classification is as observed in file “slht1Q1Um\_predictions.csv”.