**Decision Tree Classifier**

The dataset used for this classifier included a table that contained the explanatory variables - version/book-count, genre\_id, number-of-ratings, rating and the class variable 1/0. Another worksheet genre\_id (Appendix 1 – Worksheets) held the genre names that were used and the corresponding id’s. The class variables were changed to 0 or 1, representing no or yes as the algorithm worked with numeric data.

The goal was to predict if a book would be made into a movie given the above explanatory data.

In Python, the algorithm used was sourced from <https://stackabuse.com/decision-trees-in-python-with-scikit-learn/>.

**Algorithm**

The dataset was read in and stored as ‘dataset’, the data was then prepared by dividing it into attributes and labels and further dividing the resulting data into training and test data sets. This allows the algorithm to be trained on one set of data and tested on a different set of data. This was done using the sklearn library that contained a function ‘train\_test\_split’. This function allows for the specification of the percentage of the dataset to use as the test data. For this analysis this value would be varied and ran multiple times to cover a larger set of testing data, as the ‘train\_test\_split’ method randomly splits the data.

The next step was to train the decision tree algorithm to make predictions. The Scikit-Learn library contains a tree library that contains methods for various decision tree library. For the classification task, the ‘DecisonTreeClassifier’ class was used as well as the ‘fit’ method that was used to train the data and the predict method to make predictions on the test data (Appendix 2-Decision Tree Classifier). (Appendix 2-Decision Tree Classifier). Finally, the algorithm was evaluated using the metrics listed below.

**Confusion Matrix:** also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm.  Each row of the [matrix](https://en.wikipedia.org/wiki/Matrix_(mathematics)) represents the instances in a predicted class while each column represents the instances in an actual class

**Precision**, p: the number of correct positive results divided by the number of all positive results returned by the classifier).

**Recall,** r : the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive)).

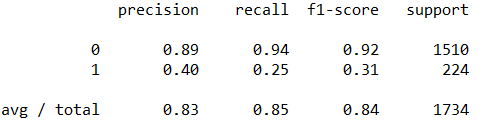
**F1-score** - a measure of a test's accuracy. It considers both the [precision](https://en.wikipedia.org/wiki/Precision_(information_retrieval)) *p* and the [recall](https://en.wikipedia.org/wiki/Recall_(information_retrieval)) *r* of the test to compute the score.

**Support:** The subset of the measurable space, number of rows for each class variable

**1st run with 20% of data as testing data(randomized)**

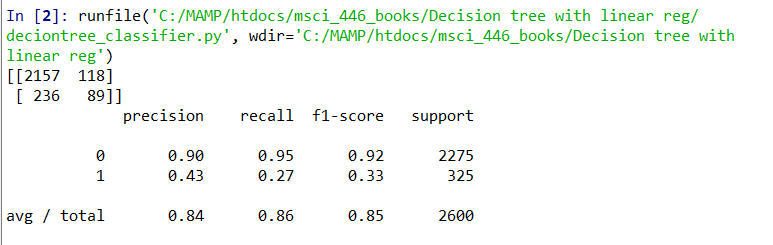


*The algorithm got 254 out of 1734 rows of data wrong, as seen from confusion matrix, which gives an 85.35% accuracy.*



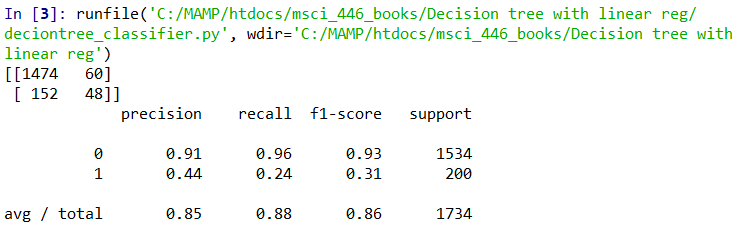
**Run 2 with test-size of 30%**

*The algorithm got 354 out of 2600 wrong, as seen from confusion matrix, this gives it an accuracy of 86.385% accuracy.*



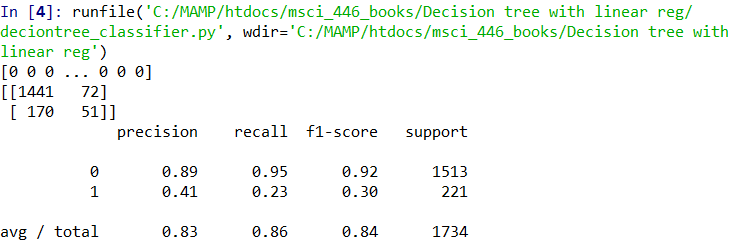
**Run 3 with 20% of data as testing data**

*The algorithm got 212 out of 1734 test data wrong, as seen from confusion matrix, giving it an accuracy of the 87.774%.*



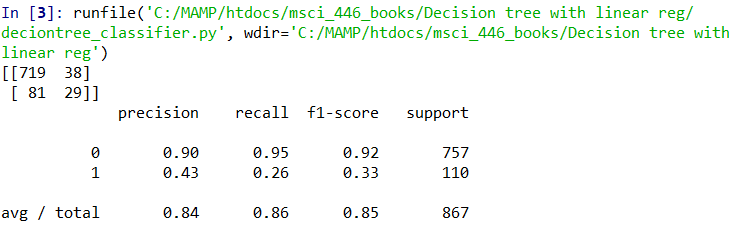
**Run 4 with 20% of data as testing data**

*The algorithm got 242 out of 1734 test data wrong, as seen from confusion matrix, giving it an accuracy of the 86.04%.*



**Run 5 with 10% of data as testing data**

*The algorithm got 119 out of 867 test data wrong, as seen from confusion matrix, giving it an accuracy of the 86.275%.*

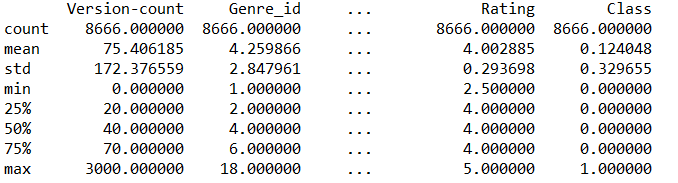


**Decision Tree for Linear Regression**

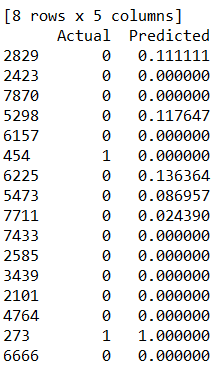
The decision tree for linear regression used the same dataset as the decision tree with classifier, it was also important that all the data being used were numeric.The goal was to predict the creation of a movie based on the explanatory variables - version/book-count, genre\_id, number-of-ratings, rating.

Algorithm

The algorithms used was similar to that used for the classification, in terms of the python libraries. The data was read in as a csv using pandas library and some the describe method produced data analysis to understand the dataset better.



The data was prepared similarly to the classifier method by dividing the data into attributes and labels and then into training and test sets. To train the data the sklearn class ‘DecisionTreeRegressor’ was used and the ‘DecissionRegressor’ and fit class and method were used to train the tree. As with the classifier the predict method was used. And finally, the predicted values and actual values were compared and then evaluated using the Scikit-Learn library ‘metrics’ package which produced following statistics:

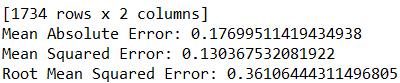


**Mean Absolute Error:** a measure of difference between two continuous variables, it is the average vertical distance between each point and the [identity line](https://en.wikipedia.org/wiki/Identity_line). MAE is also the average horizontal distance between each point and the identity line.

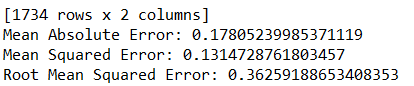
**Mean Squared Error:** assesses the quality of an [estimator](https://en.wikipedia.org/wiki/Estimator) or a predictor, this is by measuring the [average](https://en.wikipedia.org/wiki/Expected_value) of the squares of the [errors](https://en.wikipedia.org/wiki/Error_(statistics))—that is, the average squared difference between the estimated values and what is estimated

**Root Mean Squared Error:** represents the square root of the second [sample moment](https://en.wikipedia.org/wiki/Sample_moment) of the differences between predicted values and observed values or the [quadratic mean](https://en.wikipedia.org/wiki/Quadratic_mean) of these differences.

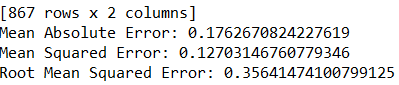
**Run 1 with 20% data as test data**



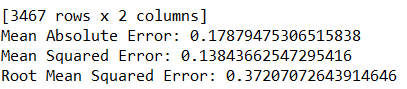
**Run 2 with 20% data as test data**



**Run 3 with 10% data as test data**

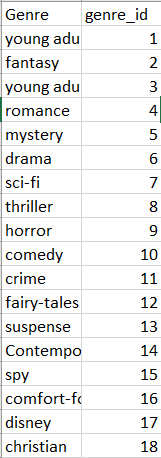


**Run 4 with 40% data as test data**



Based on the above runs, the mean absolute errors for all the runs of the algorithm produced a mean absolute error that was more than 10% of the mean of all the values in the class variable column(0.12408). This means that the algorithm did a not do a good job of predicting.

**Appendix, Data Worksheet**



*Genre Id to Genre for Decision Tree*