Capstone Project 1

In-Depth Analysis

Rate A Read With goodreads

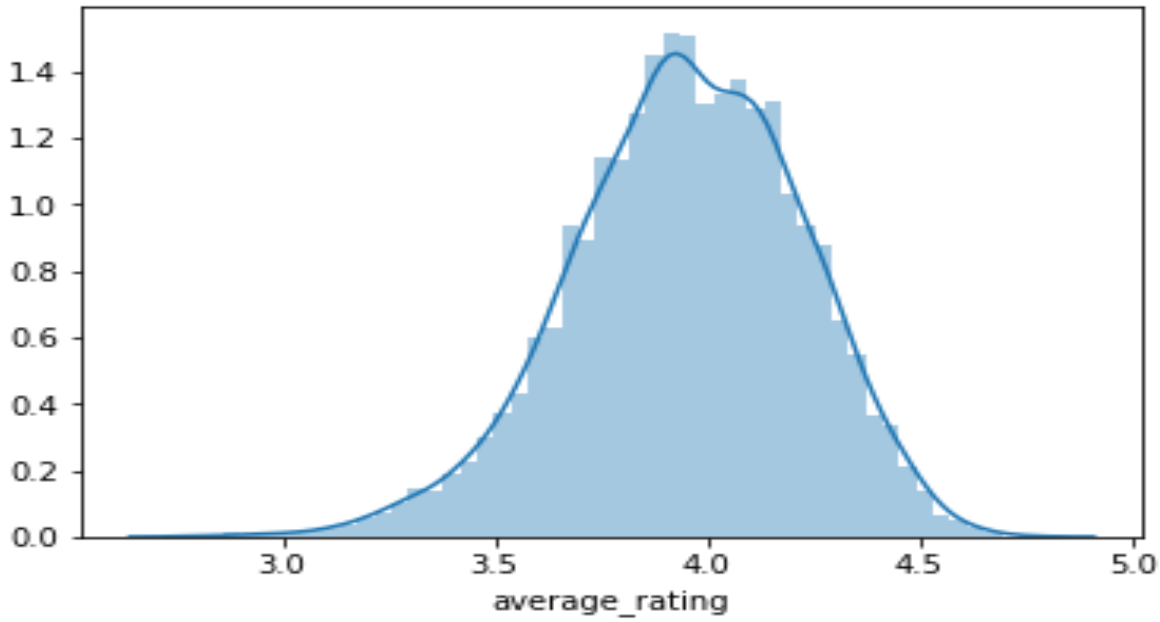
Introduction:

In this project, I have gathered Books and Author Details from goodreads using an API. Then I have explored the different features extracted from Books and Authors to determine what makes a book popular or what are the determinants in a book which earns a good rating?

This project is to predict the Average Rating of books. Since, we already have the labeled Data, this project is an example of **Supervised Learning**.

Also, we are going to predict a continuous variable and thus in this project, I have used **Regression Models** for Prediction.

Before I start fitting different Regression Models to predict the Average Rating of a Book, let's take a look at Average Rating data we have in our current Dataset.



The Average Rating data is normally distributed, which is a good start. Now, let's start fitting the Models.

Preprocessing Dataset before Fitting a Model:

1. Delete the text features or the features which are not so important for Modeling
2. Convert Boolean Columns to Integer Values
3. Convert Gender column to Integer Values

Divide Data into Training and Test set:

The Dataset is divided as Training Set and Test Set where 80% of the Data is used for Training and the rest 20% is used for the Testing Purpose.

Apply Machine Learning Algorithms:

Dummy Regressor:

A dummy Regressor or a simple baseline model is used first that makes predictions using some simple rules like always predicting the MEAN values. It is used to compare with other (real) regressors modelling done next. The score of this Model is **-0.0026.**

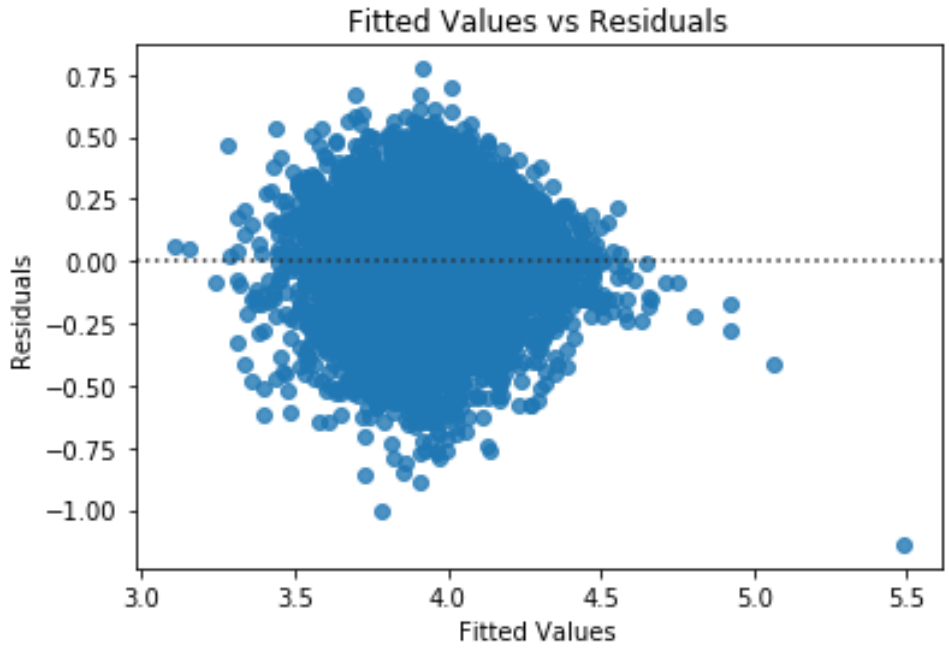
Train a Linear Regression With statsmodels

Statsmodel is a Python library designed for more statistically-oriented approaches to data analysis. It has some built in support for many of the statistical tests to check the quality of the fit and a dedicated set of plotting functions to visualize and diagnose the fit.

OLS Regression Results

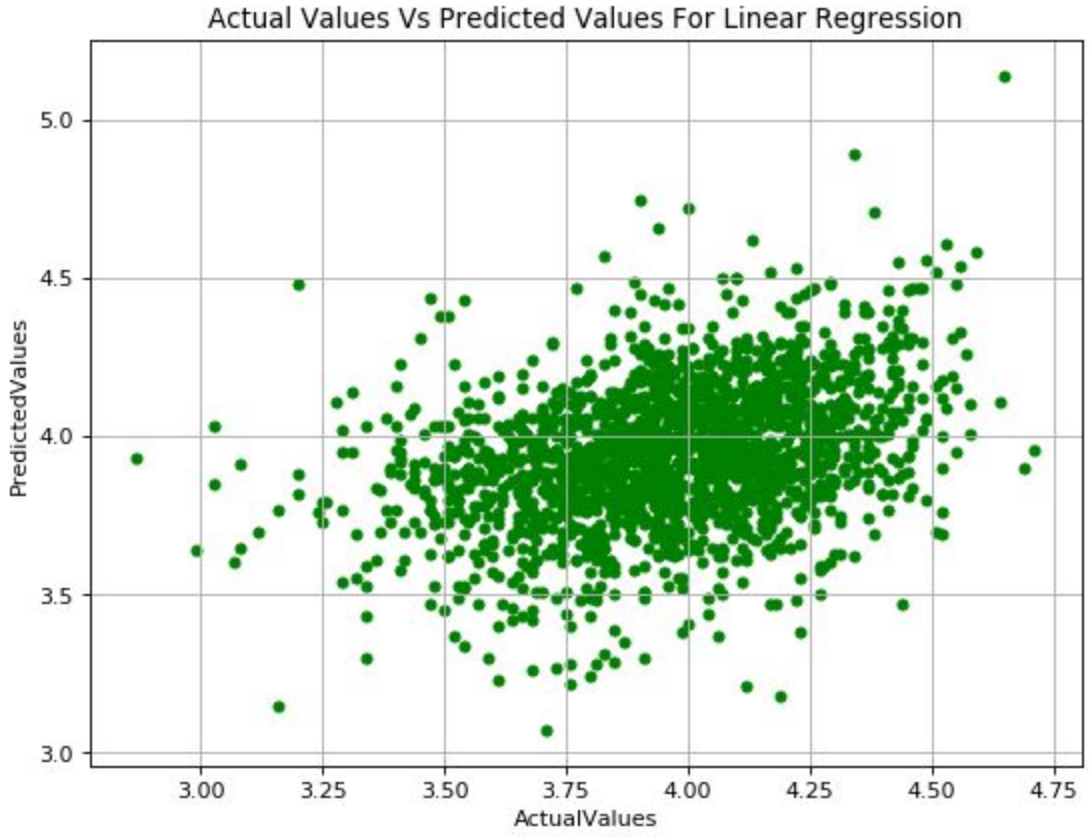
Dep. Variable:	y	R-squared:	0.441
Model:	OLS	Adj. R-squared:	0.272
Method:	Least Squares	F-statistic:	2.612
Date:	Fri, 14 Jun 2019	Prob (F-statistic):	5.05e-200
Time:	10:13:39	Log-Likelihood:	1641.5
No. Observations:	9576	AIC:	1155.
Df Residuals:	7357	BIC:	1.706e+04
Df Model:	2218		
Covariance Type:	nonrobust		

The R-Squared Value is **0.441** and the Adjusted R- Squared value is **0.27**.



Train a Linear Regression Model With scikit-learn

Scikit-learn is a powerful Python module for machine learning. In this project I have explored the sklearn.linear_model module which contains *"methods intended for regression in which the target value is expected to be a linear combination of the input variables"*.



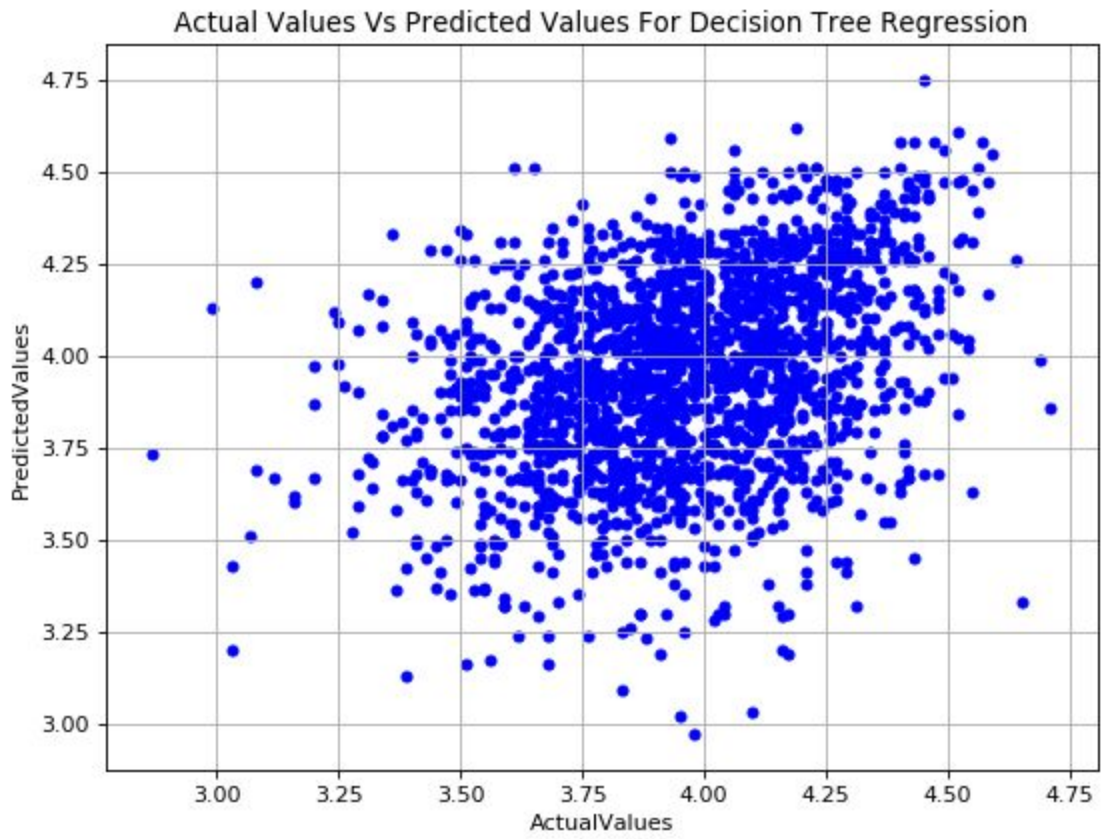
The Linear Regression Model did not perform that well and R2 score is -0.1. So, the next step is to look for some better Fit.

Train a Decision Tree Regression Model

Decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs and utility.

The branches/edges represent the result of the node and the nodes have either:

1. Conditions [Decision Nodes]
2. Result [End Nodes]



Again this model did not perform that well in our current Dataset and resulted in R2 score of -**0.300.**

Train a Random Forest Model

A Random Forest combines multiple decision trees in determining the final output rather than relying on individual decision trees. The Model Outputs the mean prediction (regression) of the individual trees.

Random Search Cross Validation

Used the below grid with 3 fold Cross Validation for Random Search:

```
{'bootstrap': [True, False],
 'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
 'max_features': ['auto', 'sqrt'],
 'min_samples_leaf': [1, 2, 4],
 'min_samples_split': [2, 5, 10],
 'n_estimators': [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]}
```

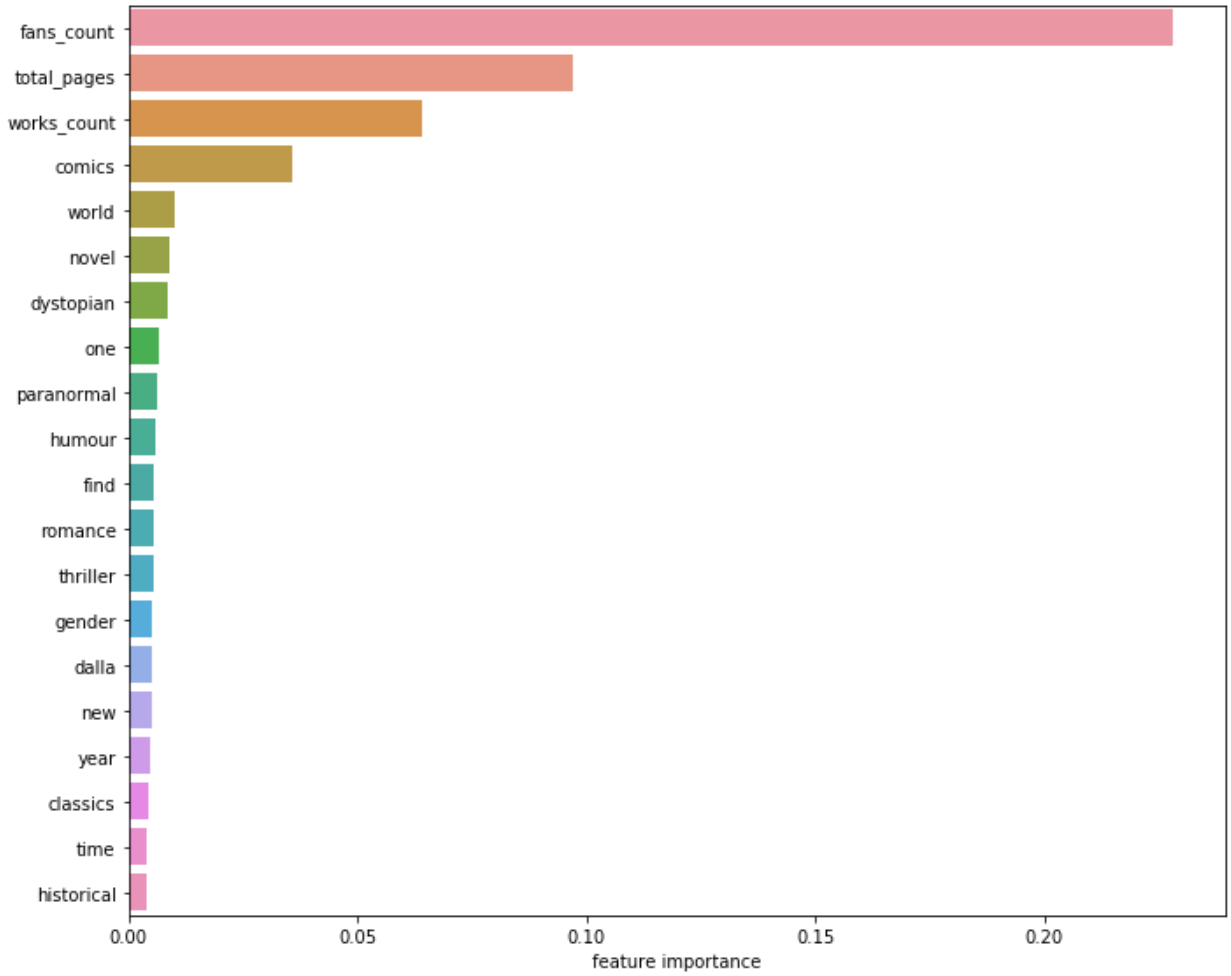
Best Parameters set found:
{'n_estimators': 900, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': None, 'bootstrap': True}

Fitted a Random Forest Model with the best parameters based on the results of Random Search.

Feature Importance as explored by RF Model

Reduced Features to explore only features with 95% importance and below are the top variables.

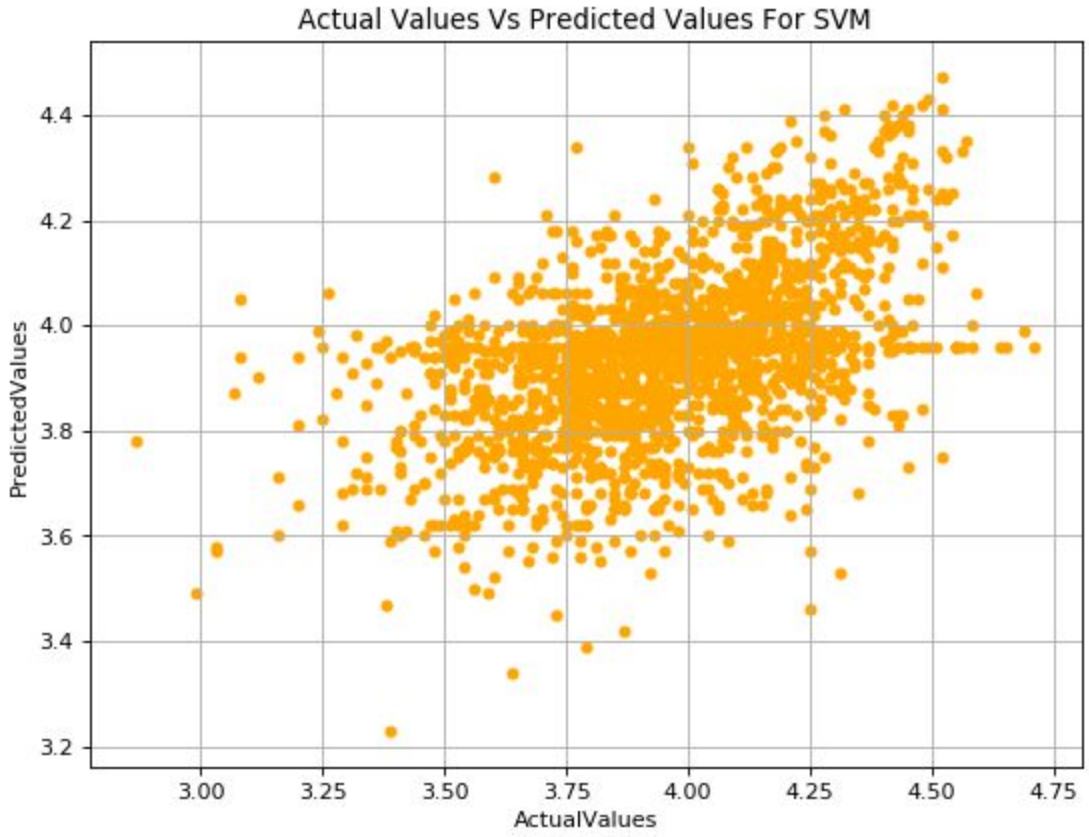
Feature Importance as explored by RF Model



Fitted a Final Random Forest Model using Best Parameters provided by Grid Search and the Reduced Features. This Model did reasonably well and the R2 score is **0.34**.

Train a Support Vector Regression Model

Support Vector Regression (SVR) is a regression algorithm, and it applies a similar technique of Support Vector Machines (SVM) for regression analysis. As we know, regression data contains continuous real numbers. To fit such type of data, the SVR model approximates the best values with a given margin called ε-tube (epsilon-tube, ε identifies a tube width) with considering the model complexity and error rate.



Compare Models And Final Results

A well-fitting regression model results in predicted values close to the observed data values. The below Metrics are used for evaluation:

1. **R2_score**: This function computes the coefficient of determination, usually denoted as R^2. It provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model. The value usually varies between 0 (worst fit) and 1 (best fit) and even it can be a negative value(because the model can be arbitrarily worse).
2. **Mean Absolute Error** (MAE) is the mean of the absolute value of the errors.
3. **Mean Squared Error** (MSE) is the mean of the squared errors. The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.
4. **Root Mean Squared Error** (RMSE) is the square root of the mean of the squared errors.

Model	R2 Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
OLS Stats Model	0.272			
Sklearn Linear Regression Model	-0.100	0.218	0.079	0.28
Decision Tree	-0.300	0.235	0.0932	0.30
Random Forest Model	0.343	0.169	0.047	0.217
Support Vector ReRegressor	0.240	0.179	0.054	0.233

In this project, I have implemented different models to predict the Average_Rating of a book and the *Random Forest Model did the best*.