A close up of a sign

Description automatically generated

**Cohort 14**

**Booksville Team**

**Elsa Abebe, Suchit Sharma, Kamal Rangavajhula, Wesenyelew Sima**

**Introduction**

E-commerce is the most popular way of selling products since the booming of e-commerce site like Amazon and Ebay. Amazon is one of the biggest cloud computing and the largest e-commerce company in America. Amazon was founded by Jeff Bezos in 1994 and its originally known for selling books through its website. Many customers think all the books sold on Amazon is actually sold by Amazon. But the reality is more than half of the books sold on Amazon is sold by third-party seller. According to an article by CNBC the Amazon CEO Jeff Bezo mention on his letter “in 2017, third-party sellers were responsible for more than half of Amazon’s sales for the first time” (Hamrick, 2019). At the beginning of the year 2019, Amazon books sale was 53.38 million (Hamrick, 2019). There are many reasons why third-party sellers choose amazon to sell books. Some of the reasons are: it is convenient to sell books on Amazon, Amazon has best seller rankings program (BSR), and Amazon has Fulfillment by Amazon Program (FBA).

Amazon made it easy for third-party sellers to sell books on their site by provides different programs .The Fulfilled by merchant (FBM) program provides sellers the option to make a listing on Amazon and to pack and ship the product themselves; Amazon Vendors(AMZ) program provides sellers the option to sell the products directly to Amazon through Amazon vendors central services; and Fulfilled by Amazon (FBA), gives the option to sellers to create a listing on Amazon, but packing and shipping the product is done by Amazon for a fee (Wood, 2019). These programs Amazon has for the third-party sellers made selling books and other products on Amazon easier.

The Best Seller Ranking (BSR) is another reason why third-party book sellers choose amazon. Amazon gives ranking to best seller books, the higher the ranking, the more books sell every day. The third, and last reason third-party sellers choose amazon is because of the Fulfilled by Amazon (FBA) program. For the third-party sellers, packing and shipping their own product might be exhausting and time consuming. The Amazon FBA program provide services like allowing sellers to ship and store their product for a fee; picking, packing and shipping seller’s products to customers; handling customer service issues like returns, exchanges and complaints on behalf of the seller; and admittance into Amazon’s Prime program, which provide faster shipment and delivery to customers (Wood, 2019). Because Amazon made selling products convenient on their site, third-party book sellers made a huge profit from selling books on Amazon.

**Hypothesis**

While selling books on amazon is convenient and profitable, there are also challenges for the third-party sellers. According to Media Corporation article most third-party sellers are fed up with Amazon (Erickson, 2018). An article by Robin Lewis “The Amazon Squeeze:You Choking Yet?” described the relationship between Amazon and its third-party sellers as “dance with the devil…The third-party vendors are between a rock and a hard place. They are both friends and enemies with Amazon.” (Lewis, 2018). The reason behind this statement is Amazon is a friend with the third-party sellers because it provides the site and customers traffic for the sellers. At the same time Amazon is an enemy to the third-party sellers because it increasing the commission and fees on the third-party sellers (Lewis, 2018). Selling product on Amazon is becoming less profitable than it used to be. *Amazon’s Sellers Forums*-the place on Amazon’s website where seller share their experience, one of the sellers shared his experience saying *“Started selling in 2002. Back then, my minimum price on a used book/CD was $7.99. I am slowly changing that to $9.99, due to increased postage costs & increased Amazon commission. My net profits have plummeted in the last 2 yr, as I’ve been slow to raise my prices.” [*[*corazonbb*](https://sellercentral.amazon.com/forums/t/what-did-selling-on-amazon-used-to-be-like-that-is-back-in-the-day/388647/42)*]*. Amazon increasing commission and fees on third-party sellers, keeping sales data to themselves (not sharing with the third-party sellers), and putting lower price on similar products to sell amazon branded products faster, made amazon third-party sellers lives difficult.

According to the posts by Amazon third-party bookseller on *Amazon’s Seller Forum*, Books sellers are facing challenges of putting the right price on books. On a post about “questions about pricing and repricing books” on Amazon’s Seller Forum, a book seller described his frustration on pricing and repricing books as “*I have been selling books on Amazon for 14 years. in 2015 I sold 40,000 books. I do Merchant Fulfillment (Not FBA). My questions are directed for those of us who are thinking long term and not just trying to sell fast and cheap. I do use a third party software for listing on Amazon. How do you price a book? do you use a different minimum price for each book? based on what? how often do you run your repricer? what is your algorithms? I am describing each book I sell and take the time to do it. I am mainly frustrated to find out that my competition is not and yet, my prices go down. too frequently (in my opinion), I sell a book for a lower price than what I think it could have been sold (for various reasons). How do we stop the race down in prices and at the same time, stay competitive and turn inventory?*” (thenextpagebooks, 2016)

We the members of Booksville team, tested a hypothesis: *one can predict the price of a book based on category (textbooks are more expensive than sci-fiction novels), edition (newer editions are worth more than obsolete editions), previous sales price, frequency of books sold, seasonality of demand and other factors.*

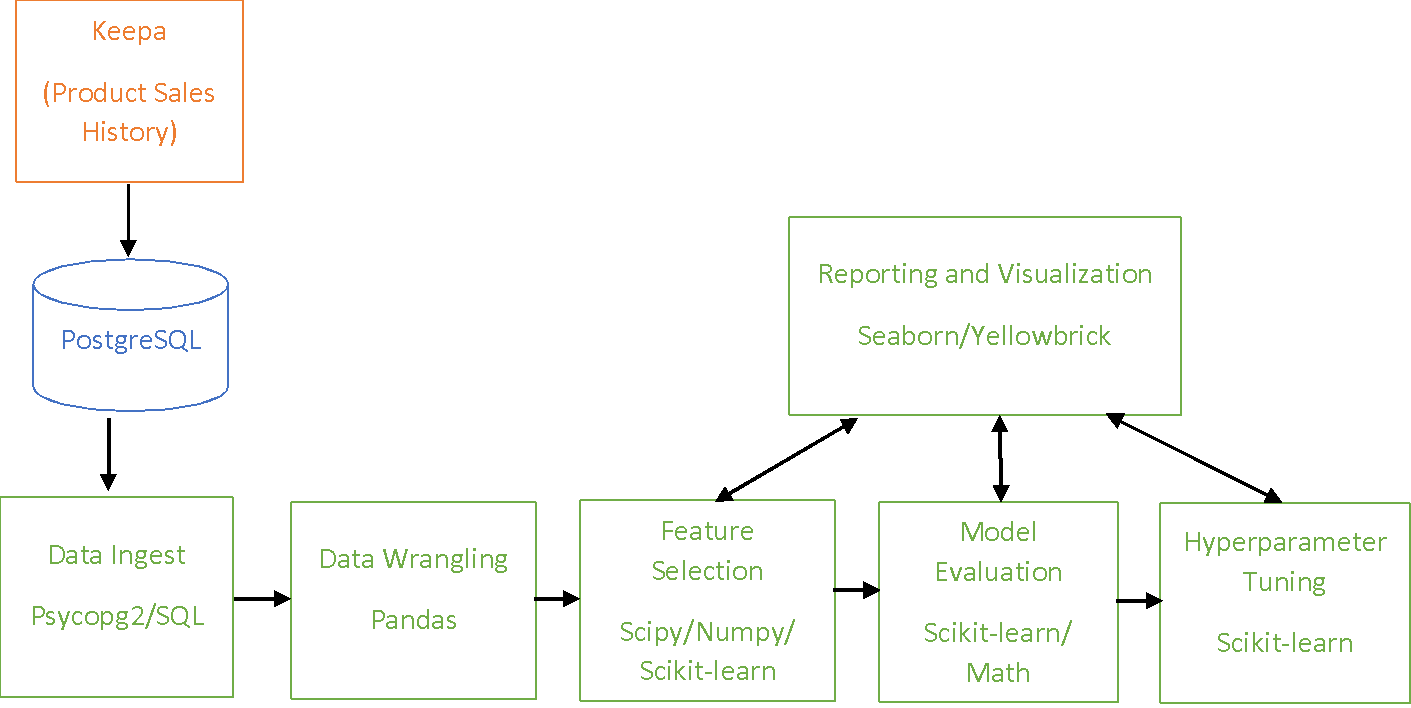
We are using data from Keepa, a tracking site that collects historical price for all products on amazon. We have obtained a record of 100,000 books from Keepa in postgres dump format and we are using the data to predict an optimal sale price for books on Amazon. Keepa is a data consolidator for the Amazon e-commerce website, the company collects all the information related to the product sold in the e-commerce platform from the ASIN code, product category, size and dimension of the product, various sellers information, price tracking information of the product. The Keepa collects the data extensively for each and every product sold on the Amazon platform.

**Application and Motivation**

Our goal is to investigate the application of supervised machine learning algorithms

towards accurately predicting the sale price of books on Amazon e-commerce site. For reducing the scope of the project, the team is mainly focusing on the books category in the Amazon platform. The data model will be based on product sales history dataset provided by Keepa which tracks sales data of every product sold on Amazon.

**Project Pipeline**



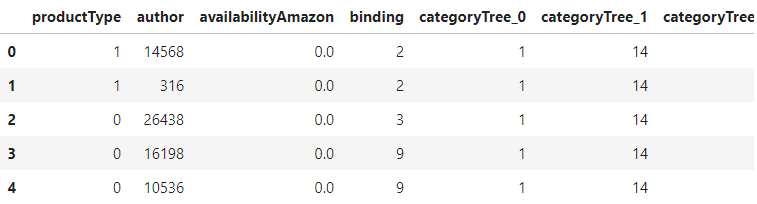
**Data Ingestion and Wrangling**

The ingestion process for the project started by retrieving the data from Keepa, we received a Postgresql dump from Keepa with 100,000 records. We restored the sql dump file into a PostgrSQL database. Then, we loaded the data into python to perform some data preprocessing steps on the data from Keepa. On our wrangling process, we examined the data for completeness and correctness to identify the intensity of missing values and resolve either by replacing or deleting the feature with the missing values. The data contain around 32 attributes with number of blank rows are more by 50% than the unique values. On our data wrangling effort we executed the following actions: removal of columns that contain no data and columns with single values because we believe the columns don’t have significant impact on determining the optimal price of the book; dropped around 20 features some of them based on yellow brick feature selection score and the others due to duplicate records, low variance, and sparse data; indexed the data with the ASIN attribute, which has unique values for each row of the data; removed details that don’t apply to the context of books; renamed generic contents to more descriptive and meaningful names to avoid confusion; and converted the categorical data into numeric format by using label encoding.

The first step of data preprocessing was to impute the missing values with the most accurate or meaningful data (whether to use median, mean, or labeling with a specific value). Setting a requirement to drop columns that doesn’t meet the minimum threshold, for example dropping columns that contains more than 50% null values was one of our focuses on this stage. On the first stage of exploratory data analysis, we examined the relationship between the various attributes and their impact on the label attribute. Then, we explored the possibility of imputing the derived attributes from the dataset and visualizing them to get more information from the dataset. On the effort of attribute examination, we looked into the possibility of creating a new attribute from the existing attributes to increase the accuracy of the model. The main focus for the project was to look at the correlation between the average price for the product in the past 30, 90, 180 days with the current price, which the machine learning model needs to predict. On this stage we also assessed the decision making on choosing which type of encoding technique needed to be applied on the columns.

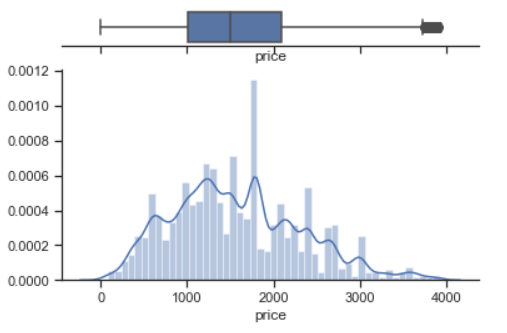
**Exploratory Data Analysis**

After the initial data wrangling, the data had 99600 records with 49 columns. For feature selection code to work, all categorical columns needed to be encoded. Using Scikit-learn LabelEncoder 15 categorical columns were encoded.



Outliers such as extreme price points or very rare categorical combinations can skew results. Executing Scikit-learn skewtest on raw data, skew of price was given as 49.71, therefore outliers had to be identified and removed for the data set.

Outliers were identified as values below 1.5 \* 25th percentile and above 1.5 \* 75th percentile of the interquartile range, also referred as Tukey’s fences. Records of these values were dropped from the dataset, resulting in the following distribution:



Next step was to identify features which don’t contribute significantly to price movement. Using scikit-learn VarianceThreshold, features with less than 20% variation were identified and dropped.

The instance was reduced to 94,981 records with 37 features.

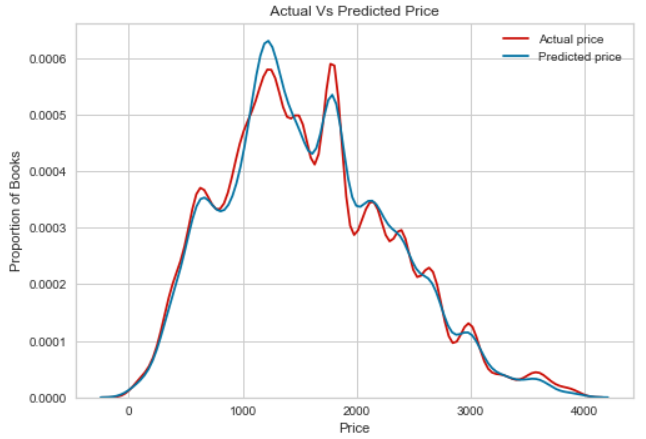
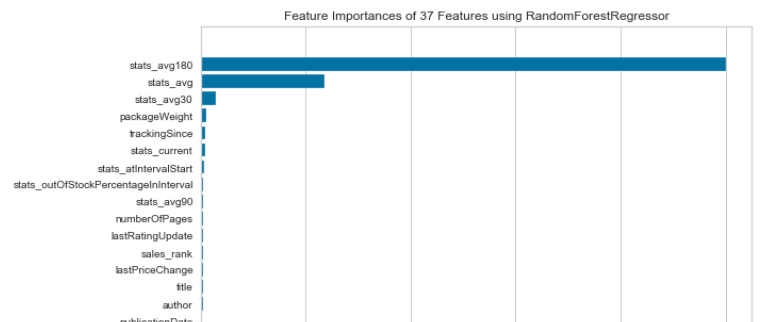
**Feature Selection**

We used Yellowbrick feature selection method for finding and selecting the most useful features and eliminate zero importance features from dataset. We executed three regression models using Yellowbrick and identified set of common significant features.

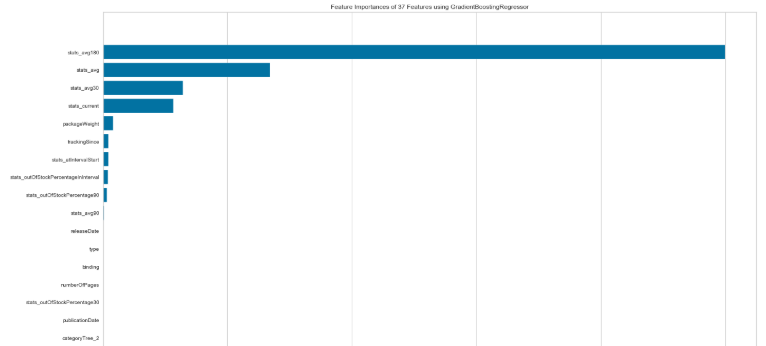
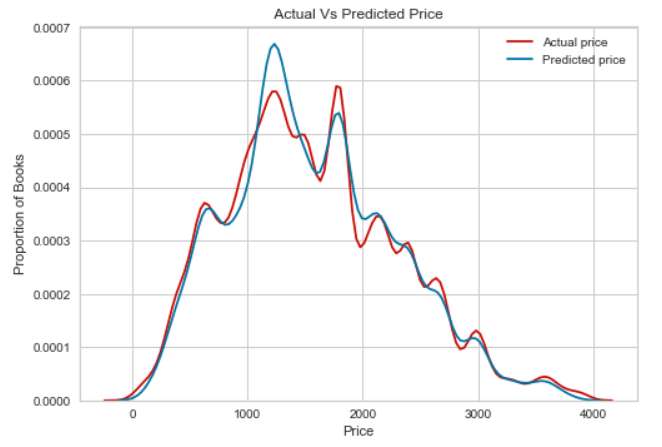
Yellowbrick FeatureImportances method calculates significance of each feature with respect to price and plots the list sorted in descending order. Used in conjunction with a Seaborn distribution plot to visualize Actual price vs Predicted Price using the significant feature list produced by FeatureImportances method, provides visual validation of Feature Selection.

Following are the visuals of FeatureImportance methods for each of the three regressors and their actual vs predicted distribution plots.

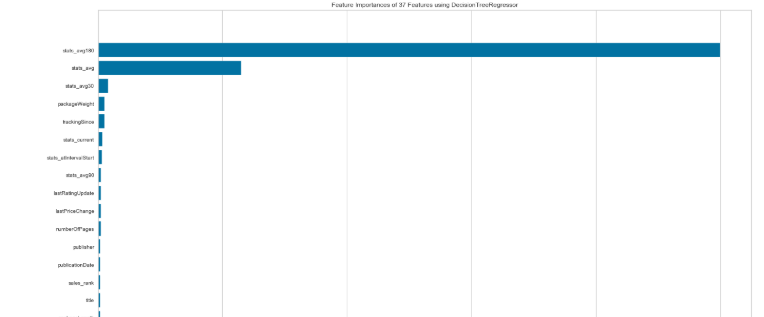
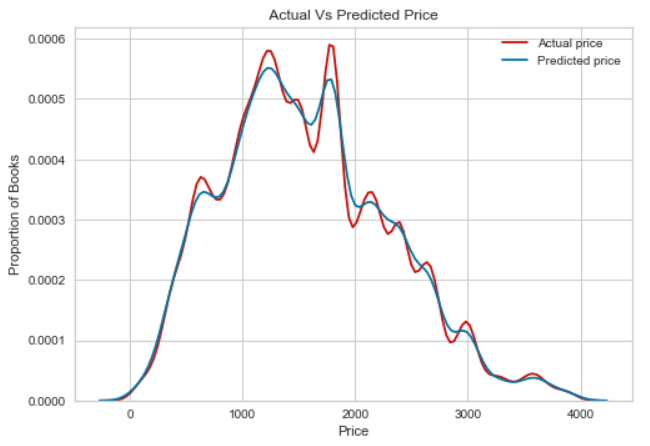
*Important Feature Selection for Random Forest Regressor*



*Important Feature Selection for Gradient Boosting Regressor*

*Important Feature Selection for Decision Tree Regressor*

Feature Selection section ends with dataset down to 94,981 records and 37 significant features.

**Model Evaluation**

For the model construction and evaluation phase we had implemented the funnel approach. Firstly, the idea was to run all the regression based algorithms on the data, and based on the metric values shortlisting the algorithms which performed better and then applying hyperparameter tuning to fine tune the model so that price of the book could be predicted better.

The only constraint while choosing a regression algorithms is the MemoryError or running/processing time of the algorithm.

Using Scikit-learn, the following models were tested for Accuracy Score and MSE:

|  |  |  |
| --- | --- | --- |
| Model | Accuracy Score | RMSE (Root Mean Squared Error) |
| Gradient Boost | 0.9367 | 186.892 |
| Random Forest | 0.9337 | 191.185 |
| Decision Tree | 0.8823 | 254.817 |
| Linear Regression | 0.3394 | 603.907 |
| MLP | negative | 3.666 M |
| Ridge CV | negative | Very High |
| LassoLars | 0.6205 | 457.734 |
| Lasso | 0.8548 | 283.051 |
| Elastic Search | 0.8502 | 287.564 |
| Bayesian Ridge | negative | 62.41 K |
| Ransac | negative | 2.89 M |

For the next steps, based on the output of the algorithms, top 4 performing which are Random Forest Regressor, Decision Tree Regressor, Gradient Boosting Regressor, and Lasso Regressor are shortlisted. Then by using the Yellow brick’s residual plot package the residual plots and the distribution of the errors for the training and testing data can be visualized. The following visualizations illustrates the residual plots for the four top performing algorithms.

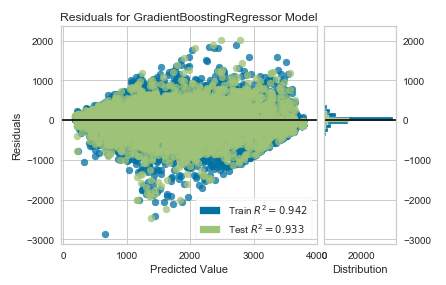


Fig a: The residual plots for Gradient Boosting Regressor Model

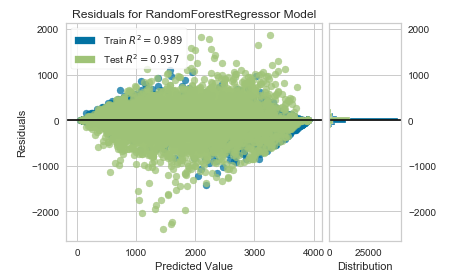


Fig b: The residual plot for Random Forest Regressor Model

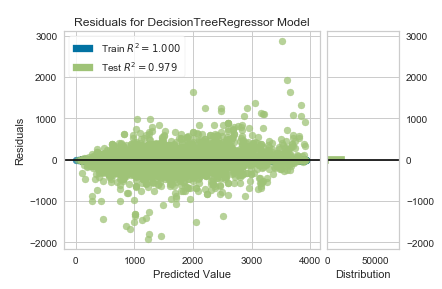


Fig c: The residual plot for the Decision Tree Regressor Model

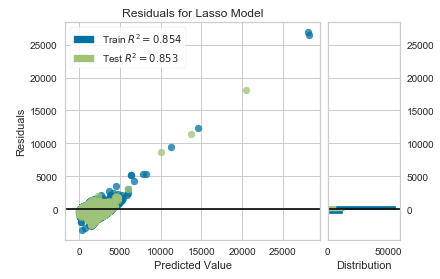


Fig d: The residual plot for Lasso Model

From the illustrations it is clear that even though the Lasso model has a considerable accuracy score, the residual plot and distribution shows other wise and the linear model is not suitable for predicting the price value.

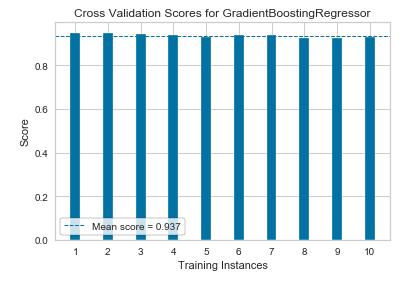


Fig 1(a): The Cross Validation Scores for Gradient Boosting Regressor algorithm.

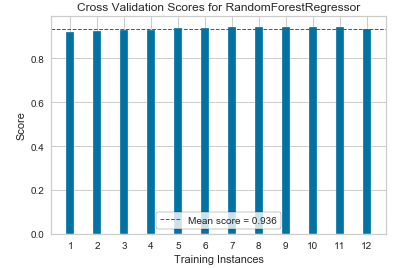


Fig 1(b): The Cross Validation Scores for Random Forest Regressor algorithm.

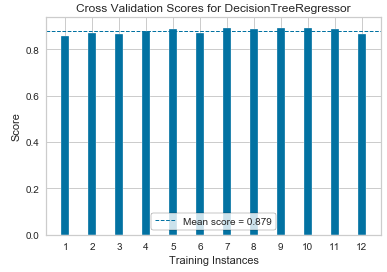


Fig 1(c): The Cross Validation Scores for Decision Tree Regressor algorithm.

Then from the residual plots, only three algorithms which are able to better fit the dataset which are decision trees, random forests, and gradient boosting regressor are considered for the next step of applying Stratified KFold which helps to train the model better i.e. by training on different sample sets of the data, and helps to find whether the model is overfitting the data or not. The above visualizations show the StratifiedKFold for the three algorithms with 12 different sample datasets.

Coming to the most important step in the hyperparameter tuning which GridSearchCV here the various range of values are given as input for each attribute in the algorithm. For example for Gradient Boosting Regressor the values for the alphas, sample split, maximum depth, learning rate along with number of folds are given as input for the GridSearchCV and after running all the possible combinations and looking for a best value which further increase the accuracy score can be found by using grid.best estimator code snippet. Which gives the best input combination for the Gradient Boosting Algorithm and the same is repeated for the Random Forest Regressor and Decision Tree Regression algorithms.

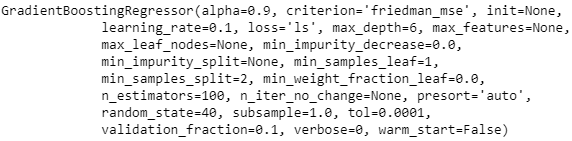


Fig 2(a): The GridSearchCV for the Gradient Boosting Regressor Algorithm.

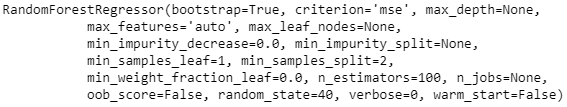


Fig 2(b): The GridSearchCV for the Random Forest Regressor Algorithm.

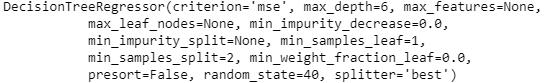


Fig 2(c): The GridSearchCV for the Decision Tree Regressor Algorithm.

In the last step, the output value from the GridSearchCV is given as an input for the top three algorithms, and the accuracy score, root mean square error values are measured. Certainly, there is an increase in the performance of the model after applying the grid search values as an input for the model.

Furthermore, to further reduce the dimensionality of the data, and to improve the performance of the model Standardization and principal component analysis had also been applied on the dataset. Interestingly, after applying the standardization on the dataset, and repeating the funnel strategy there is actually an improvement in the performance of the model, as it a prerequisite to apply standardization before applying Principal Component Analysis that is the reason why the data was standardized so that the entire data is in one standard deviation away from the mean value which increases the central tendency of the data and helps to neutralize the outlier values and gives equal importance to each and every parameter in the model. Next, the principal component analysis is applied on the data and the output contains only 8 columns instead of 37 columns which in turn reduced the accuracy score and increased the error rate. Insight after applying the dimensionality reduction algorithm is there is a trade off between the accuracy score and the computational resources required for the model.

**Results**

From the funnel strategy, the top performing algorithms after applying hyperparameter tuning are as follows:

Random Forest Regressor: 0.9404 Accuracy Score and RMSE value of 181.3

Gradient Boosting Regressor: 0.9394 and RMSE value of 182.84

Decision Tree Regressor: 0.9277 and RMSE value of 199.66

Interestingly, Gradient Boosting Algorithm had performed better than Random Forest Regression until the Hyperparameter Tuning. Once, fine tuning was done on the models the Random Forest outperformed other algorithms.

**Conclusion**

With 94% accuracy, Booksville project proves that it is possible to predict future price of a book if a healthy sale history is available. As expected, as the sale history deprecates so does the accuracy of prediction.

The most critical element of any data science project is data, without data there is no progress possible. While this applies to Booksville also, this project is significantly dependent on a subclass of data aka statistical data. Statistical data is averaged sale history for different methods of sale – Amazon, 3rd party, new / used / trade-in, previous editions, affiliate – part of series, etc. This class of features is Amazon’s Ace in capturing historical knowledge in their dataset. As long as affiliate or edition statistical history is available, Booksville can even reliably predict sale price of a new book.

**References**

David Hamrick (2019, February 2). How To Sell Books on Amazon:Tips & Secrets for 2019*.* Retrieved June 06, 2019, from<https://www.junglescout.com/blog/how-to-sell-books-on-amazon/>

KJ Erickson (2018, July 3). In Their Own Words: Why Sellers Are Fed Up With Amazon. Retrieved June 5, 2019, from<https://medium.com/public-market/in-their-own-words-why-sellers-are-fed-up-with-amazon-e97da44f7f18>

Robin Lewis (2018, February 1). The Amazon Squeeze: You Choking Yet? Retrieved June 5, 2019, from<https://www.therobinreport.com/the-amazon-squeeze-you-choking-yet/>

Meredith Wood(2019, January 12). How to Sell on Amazon: The Ultimate Guide for New Amazon Sellers. Retrieved June 5, 2019 from <https://www.fundera.com/blog/selling-on-amazon>

Amazon Services Sellers Forums (2018).<https://sellercentral.amazon.com/forums/t/what-did-selling-on-amazon-used-to-be-like-that-is-back-in-the-day/388647>

Amazon Services Sellers Forums (2016).<https://sellercentral.amazon.com/forums/t/questions-about-pricing-and-repricing-books/108095>