Title: **Review Rating Prediction**

Name: Bhushan Suryawanshi

Date: 2020-10-10

### **1. Introduction** –

With the e-commerce boom, there are multiple products available to consumers. While deciding the best product as well as recommending products for the user e-commerce enterprises rely on the product review. Having these product review data in good quality will also help analytic teams to come up with a better model. I am using amazon product reviews for the analysis. I will be using existing reviews to find the review rating based on the text. The review rating ranges from 1 to 5, where 1 is the lowest, and 5 is the max rating. For example, if the review contains the text “Good product to be used.” The model will try to see what can be the review rating for this review. This will help improve the quality of data by detecting review ratings based on review text if it is missing. As well as ensure that the review ratings are in compliant with review text.

**2. Data**

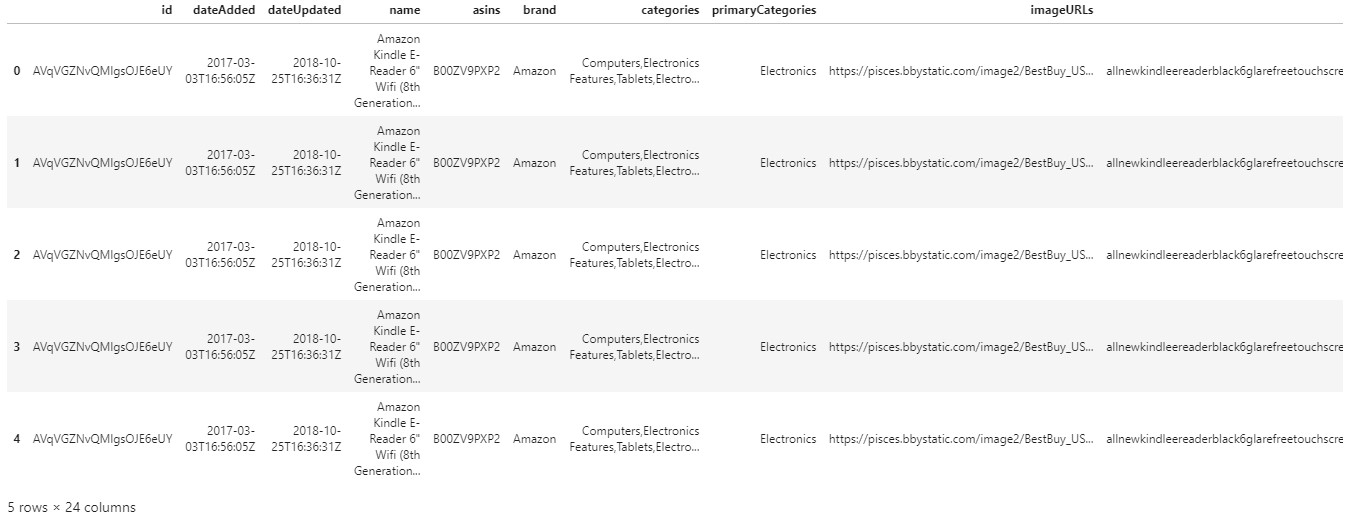
The dataset is a list of consumer reviews for multiple products. There are a lot of details in this dataset but I will be using - name, primaryCategories, manufacturer, reviews.doRecommend, reviews.rating, and reviews.text for my analysis.

Data Source Link -

https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products

### **3. Graphical Analysis**

3.1 Lets collect the data and load the dataframe to be used for our analysis.



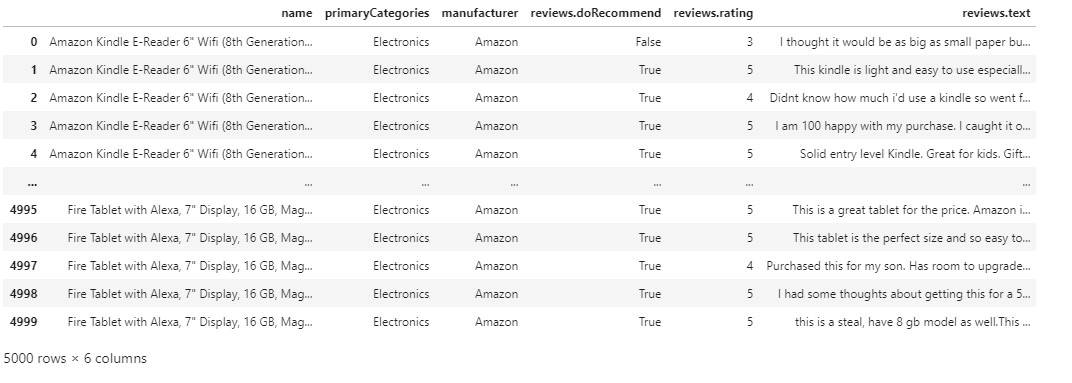
3.2 I am checking the dimensions of my dataframe. So looks like we have 5000 observations having 24 features.



Too many columns hence we will look at one row for better understanding.



There are 24 columns and for our analysis we don't need url or product ID type data. Lets select only required columns.



Renaming the column names for better access.

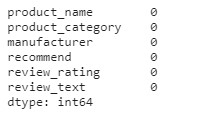


Since we have selected only required columns the dimension of dataframe will change. Lets check it.



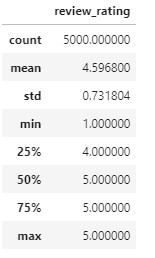
We can see there are 5000 rows and 6 columns.

3.3 Verify if any missing values.

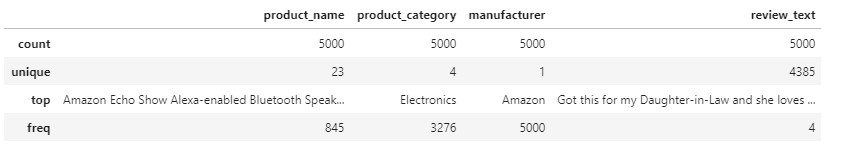


So we can see that there are no null values in any columns.

Describe data -

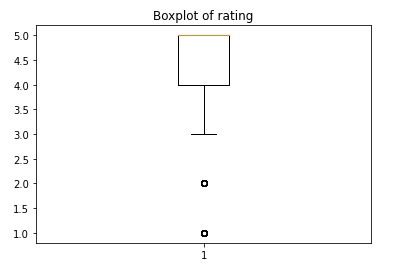


Review rating is the only numeric data in our dataset. Max is 5 and min is 1 which reconfirms there are null or zero values.



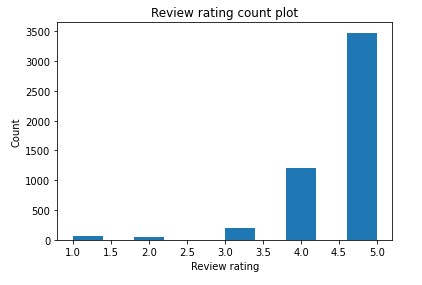
Data summary shows that, there are 23 products from 4 unique categories for one manufacturer.

3.4 Use box plot for review rating.



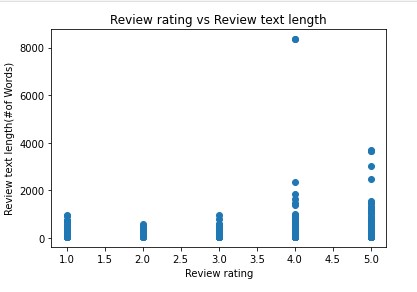
Boxplot shows me that the data is skewed towards review rating 4 and 5. Lets look at the count plot and see the details for better understanding.

3.5 Count plot for each rating.



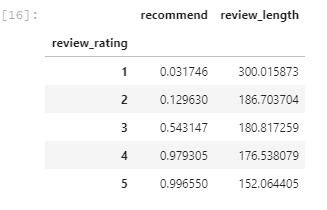
Yes, as shown in boxplot the cont plot also confirms that more than 60% reviews are rated with rating of 5.

Review text length vs Review rating graph below shows that the reviews with rating 5 are also having longer reviews. Can I say its because customer liked the product and got excited to talk more about it?

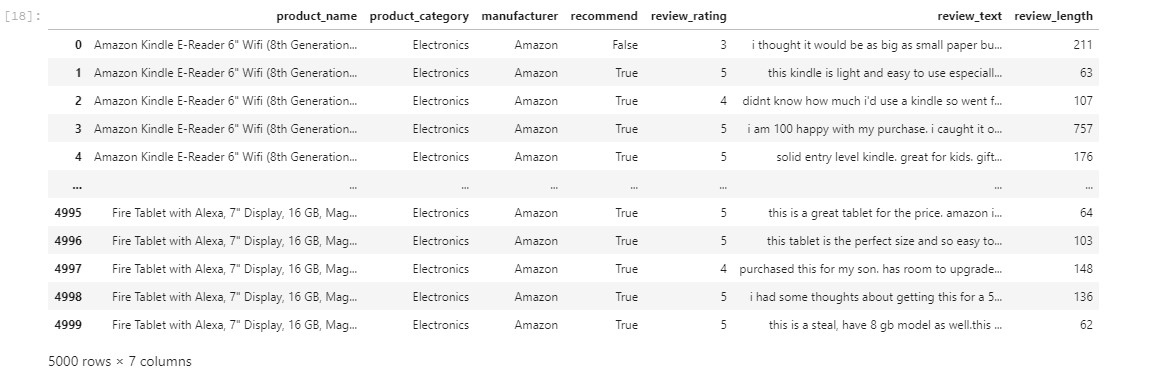


### 4. Feature extraction

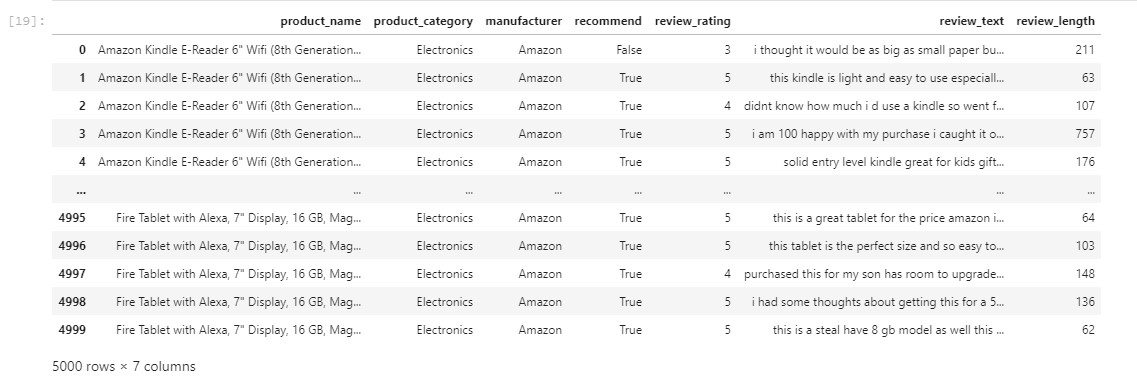
4.1 Lets check the mean of text length per review rating to understand variation.



4.2 As part of preprocessing the text data lower case the review text.



4.3 Remove punctuations.



4.4 Take out stop-words as those are not holding any important information.



4.5 Apply porter-stemmer to the text data.



4.6 Split the pre-processed data into training and test set.

4.7 Apply TF-IDF vectorizer to extract feature matrix of text data.



4.7.1 Lets check TF-IDF vocabulary



4.8 Applying additional feature extraction technique (count-vectorizer) which can be used for comparing model performance.



4.8.1 Checking count-vectorizer vocabulary

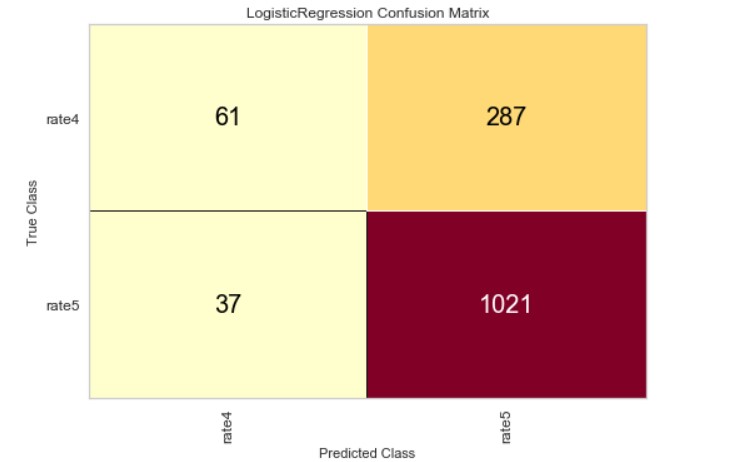


### **5. Model Evaluation**

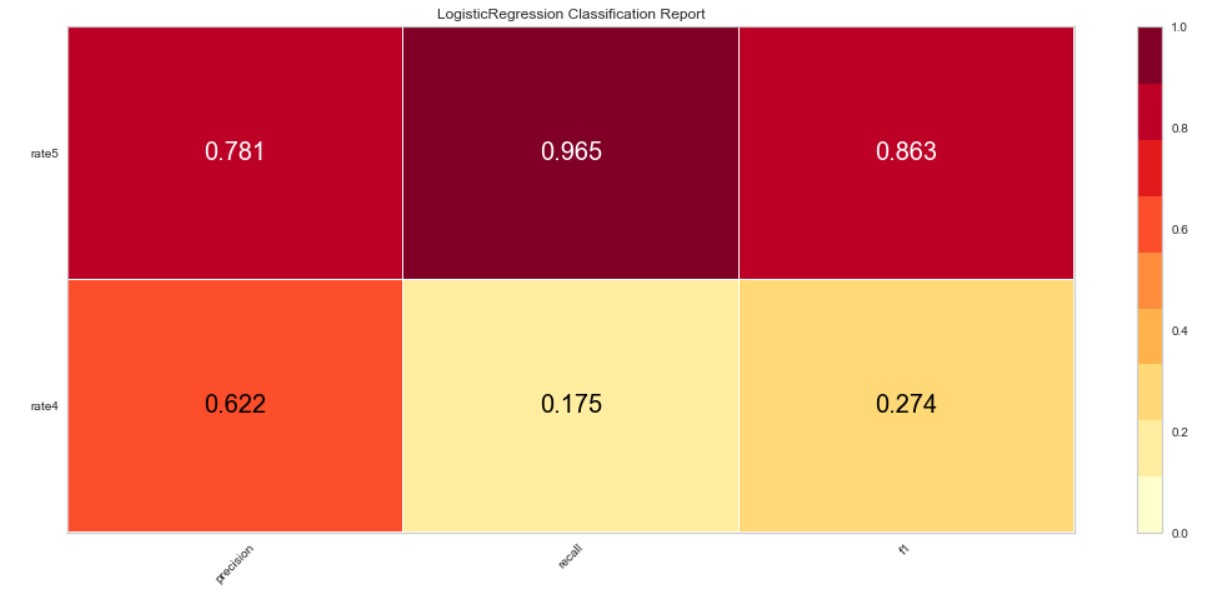
5.1 Using Logistic regression for classification model.

I jumped to use the existing model from the Titanic study. However, using it for my task was not easy and there was too many changes in terms of how features are managed. But it was great learning.

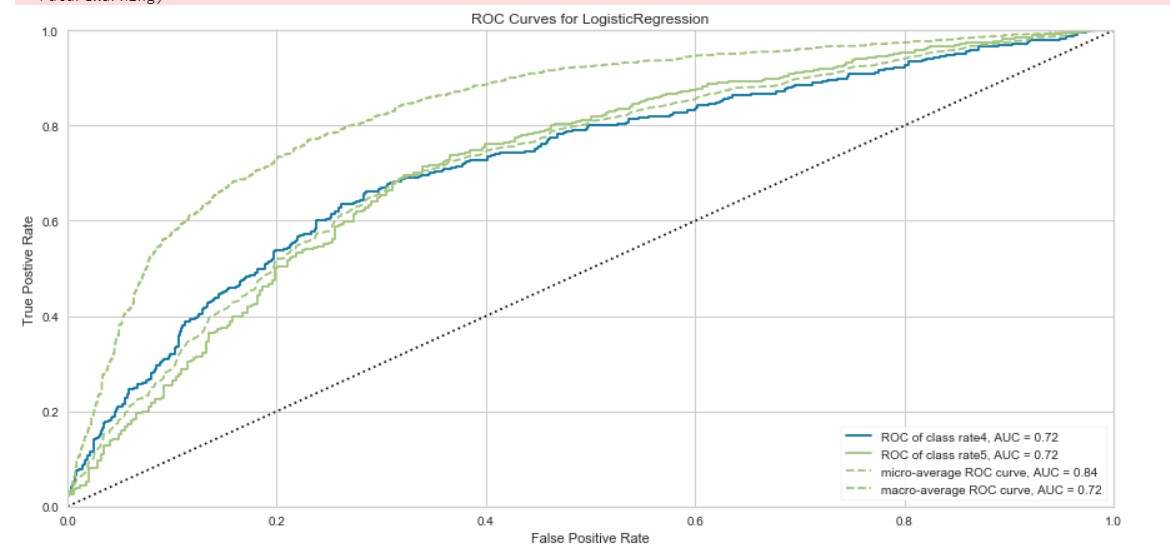
5.2 Visualize confusion matrix with model performance.



5.3 Checking what are Precision, Recall and F1 score for the model.



5.4 ROC and AUC analysis.



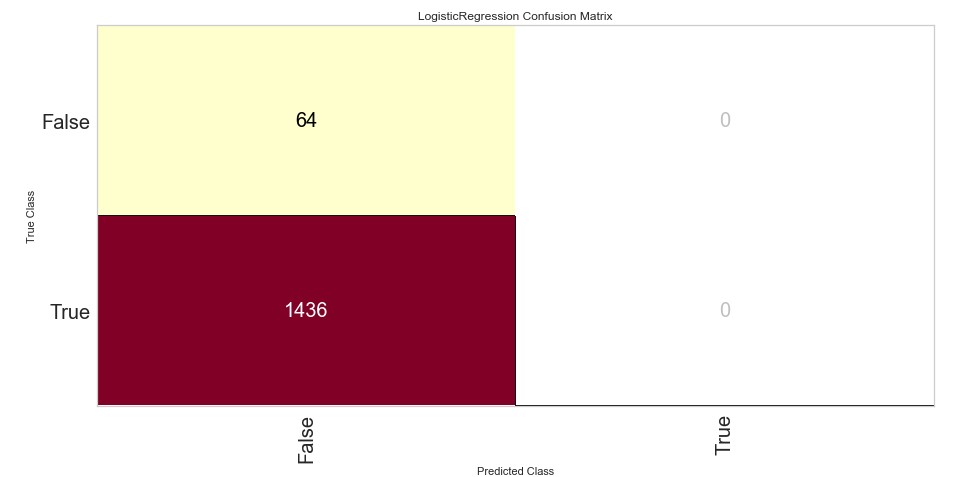
**6. Additional Models**

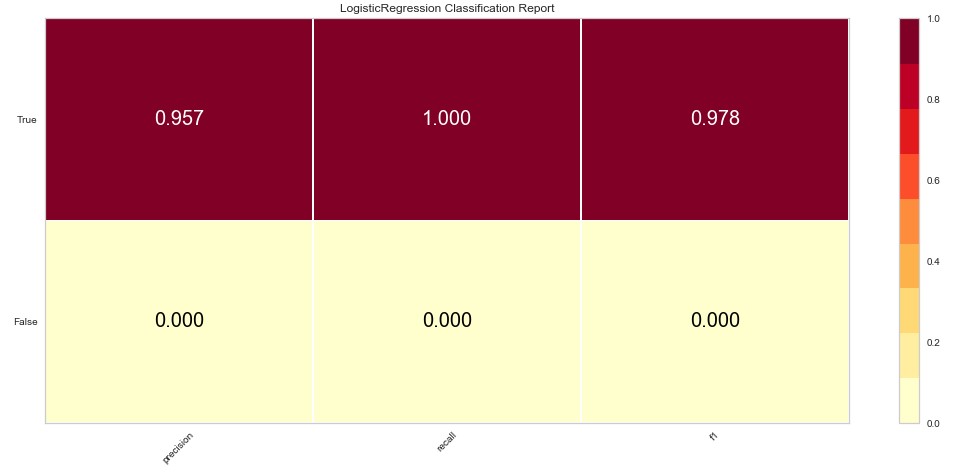
This week I tried to use different models for the same data and found that using Multinomial Naive Bayes is producing better result compared to other models I tried.

Also in one of the review response it was recommended to use "review text" and predict “review recommendation” (Recommended product – True or False). I tried to implement this using logistic regression, Multinomial naive Bayes and Decision tree.

Here are results I received with different models -

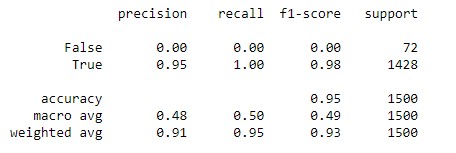
6.1 Predict if product is recommended based on the review text using Logistic regression



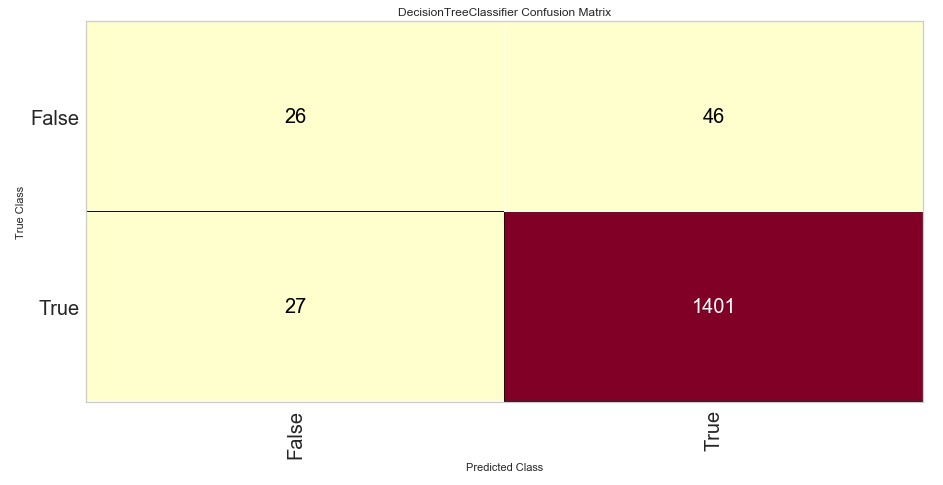


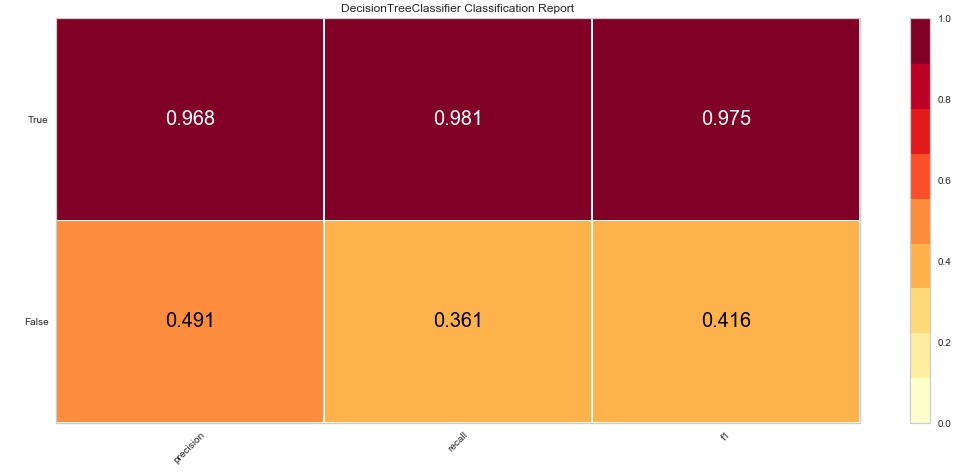
6.2 Predict if product is recommended based on the review text using Multinomial Naive Bayes



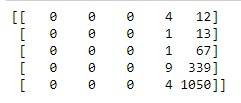


6.3 Predict if product is recommended based on the review text using Decision Tree





6.4 Predict review rating based on review text using Multinomial naive Bayes.



### 

### **7. Conclusion**

When I first used logistic regression and got higher accuracy it was like a winning moment. However, based on different suggestions on the previous milestone I went ahead and tried other models to see the difference. It clearly shows some of them improved accuracy and others did not. For the problem where I was trying to predict the “Recommend” feature based on text, a Decision tree was the best to use. I can see that one model cannot give you a real picture and you have to try out multiple of them as part of your model evaluation and selection.

# I do agree that there is more work that needs to be done to utilize the pipeline for evaluation and selection and that is what my goal is for a future project. A lot of learning from this project.

### **8**. Reference

1. https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products

2. Applied Text Analysis with Python – by Bengfort, Bilbro & Ojeda.