CS584 – MACHINE LEARNING

FALL 2016

LAPTOP PRICE PREDICTION

Group Members:

Chris Mathew Dani(A20372828)

Mabin Mariam Joseph(A20366724)

Table of Contents

Laptop Price Prediction

Group Members: Chris Mathew Dani, Mabin Mariam Joseph

Task

People often buy laptops by first fixing a given price range, but given a set of required specifications, it can be difficult to find out what the average price of a laptop is, or how the price varies with the specifications. Given the large number of laptops each having a good number of specification differences, the laptop selection problem causes difficulty with a lot of people.

Using a data that contains the list of devices, it's specifications and prices, a regression model can be created, which can predict the price for a given set of specifications. Also the contribution of each specification can be determined enabling a person to buy a laptop that fits best (budget and specifications wise). The prediction model would benefit in :

- Predicting the price of a yet-to-be launched product.
- Assisting a buyer in choosing the most suitable laptop based on required specifications.

This is interesting because this seems to be a problem that a lot of people I know have faced from time to time. To be able to set a budget for a laptop knowing the specifications they need, would help people make a good decision. Also, given a set of specifications, what would be the best laptop for a particular budget (best bang for the buck).

Dataset

The dataset is a list of all the laptops along with their specifications which was collected from Bestbuy website.

Number of data instances in the original dataset : 748

Number of features in the original dataset : 15

Number of data instances in the cleaned data set : 748

Number of features in the cleaned dataset : 64

Data source

The listed laptops from BestBuy website was scraped using python web scraping libraries BeautifulSoup and urllib. No additional data was collected and no manual labeling was performed.

Target variable

The target variable is the '**retail price**' of the laptop.

Features

The input features are the specifications of the laptop.

Features:	OS	Pro Speed	RAM	Screen Size	Touch Screen	Weight	Ratings
No. of unique values	7	31	9	17	2	193	31

Features:	OEM	HDD	SSD	Convertible	Refurbished	Pro Brand	Screen Resolution	Graphics
No. of unique values	14	12	13	2	2	6	18	6

The features are described in detail below :
OS: Operating System of the laptop (CATEGORICAL)

Pro Speed: Processor speed of the laptop measured in Gigahertz (CONTINUOUS)

RAM: Memory of the laptop (CONTINUOUS)

Screen Size: The screen size of the laptop measured in inches (CONTINUOUS)

Touch Screen: The touchscreen capability of the laptop (CATEGORICAL)

Weight: These are the weights of the laptop (CONTINUOUS)

Ratings: These are rating provided by laptop users on the BestBuy website between 0 and 5 (CONTINUOUS)

OEM: (Official Equipment Manufacturer) The manufacturer of the laptop (CATEGORICAL)

HDD: The hard disk capacity of the laptop (CONTINUOUS)

SSD: The solid state drive capacity of the laptop (CONTINUOUS)

Convertible: Whether the laptop can be converted to a tablet (CATEGORICAL)

Refurbished: Whether the laptop is refurbished or not (CATEGORICAL)

Pro Brand: The processor brand used in the laptop (CATEGORICAL)

Screen Resolution: The screen resolution of the laptop (CATEGORICAL)

Graphics: Graphic processor manufacturer (CATEGORICAL)

Data size

There are 748 data instances.

Preprocessing

- In order to avoid losing any data because of the small size, the dataset was partially cleaned by putting in missing values wherever possible.
- The dataset was then checked to see if any further missing values were there and were filled with 0.
- Categorical features were vectorized using DictVectorizer( )

FEATURE LIST AFTER VECTORIZATION

'Retail Price', 'Pro Speed', 'RAM', 'Screen Size', 'Weight', 'Ratings','HDD', 'SSD', 'Convertible=No', 'Convertible=Yes', 'Graphics=AMD','Graphics=ARM', 'Graphics=Intel', 'Graphics=NVIDIA', 'Graphics=PowerVR','Graphics=UMA', 'OEM=Acer', 'OEM=Alienware', 'OEM=Apple', 'OEM=Asus','OEM=Compaq', 'OEM=CybertronPC', 'OEM=Dell', 'OEM=HP', 'OEM=Lenovo','OEM=MSI', 'OEM=Microsoft', 'OEM=Samsung', 'OEM=Toshiba', 'OEM=type 0','OS=Chrome OS', 'OS=MacOS','OS=Windows', 'OS=Windows 10','OS=Windows 7', 'OS=Windows 8.1', 'OS=Windows7', 'Pro Brand=AMD','Pro Brand=Intel', 'Pro Brand=MediaTek', 'Pro Brand=NVIDIA','Pro Brand=Rockchip', 'Pro Brand=type 0', 'Refurbished=No','Refurbished=Yes', 'Screen Resolution=1280 x 800 (WXGA)','Screen Resolution=1360 x 768 (HD)','Screen Resolution=1366 x 768 (HD)','Screen Resolution=1440 x 900 (WXGA+)','Screen Resolution=1600 x 900 (HD+)','Screen Resolution=1920 x 1080 (Full HD)','Screen Resolution=1920 x 1280', 'Screen Resolution=2160 x 1440','Screen Resolution=2160 x 1440 (Full HD+)','Screen Resolution=2304 x 1440', 'Screen Resolution=2560 x 1440 (QHD)', 'Screen Resolution=2560 x 1600 (Retina)','Screen Resolution=2736 x 1824','Screen Resolution=2880 x 1800 (Retina)','Screen Resolution=3200 x 1800 (quad HD+)','Screen Resolution=3840 x 2160 (UHD)','Screen Resolution=3840 x 2560 (UHD)', 'Screen Resolution=type 0','Touch Screen=No', 'Touch Screen=Yes'

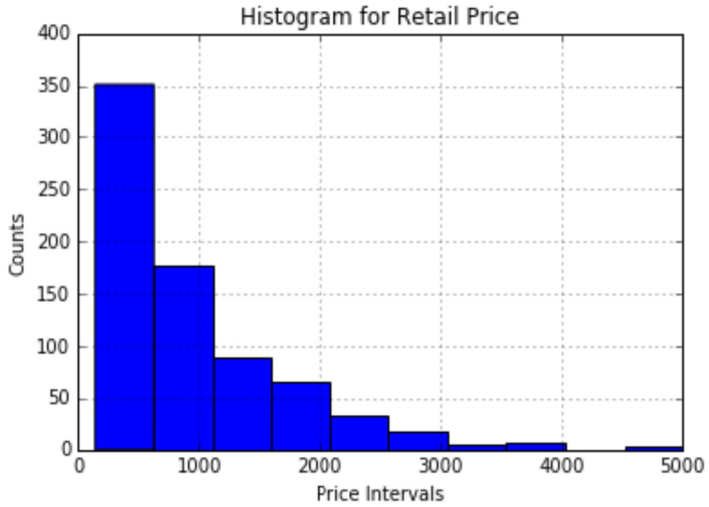
- Continuous features were scaled using preprocessing.scale( ) function.

Visualization

Target

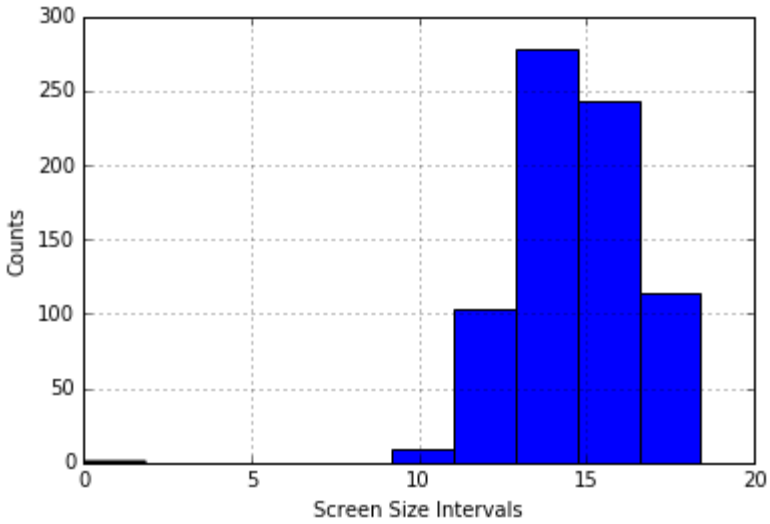
Mean of 'Retail Price': 920.79

Variance of 'Retail Price': 618616.36



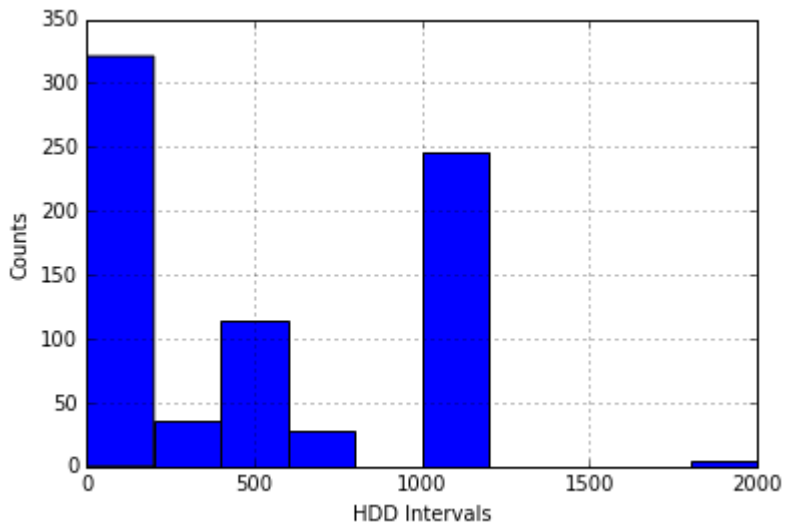
Features

Screen Size:



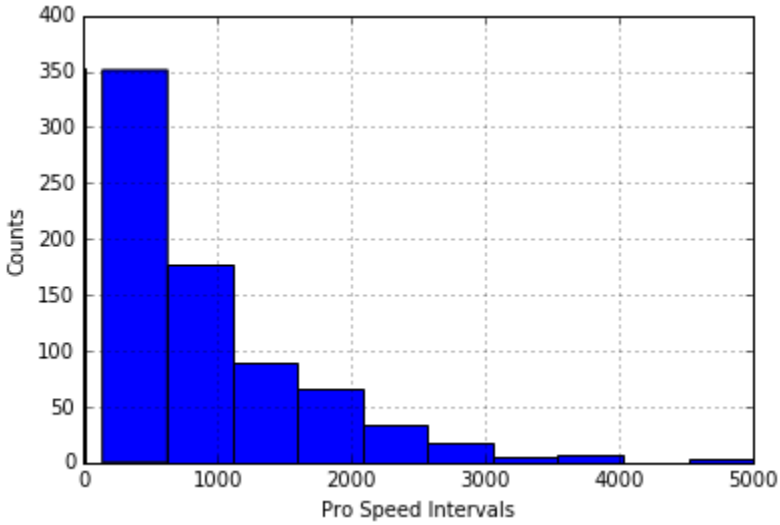
The variance of Screen Size 3.23
The mean of Screen Size 14.60

HDD:



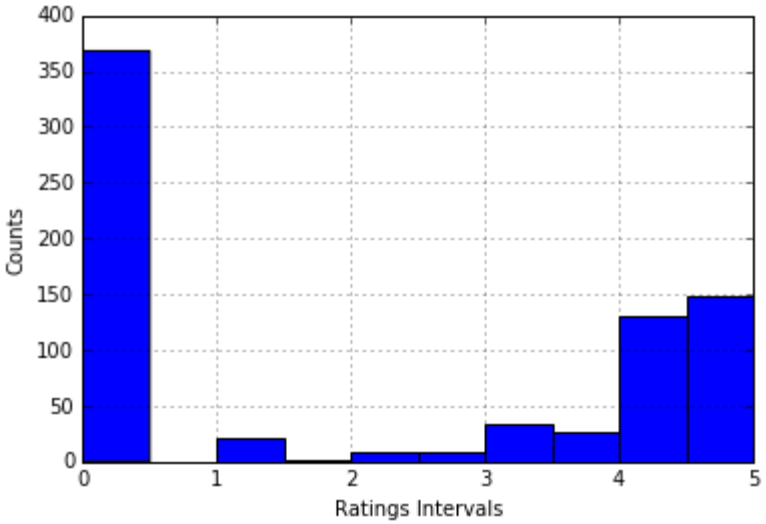
The variance of HDD 202627.63
The mean of HDD 461.61

Processor speed:



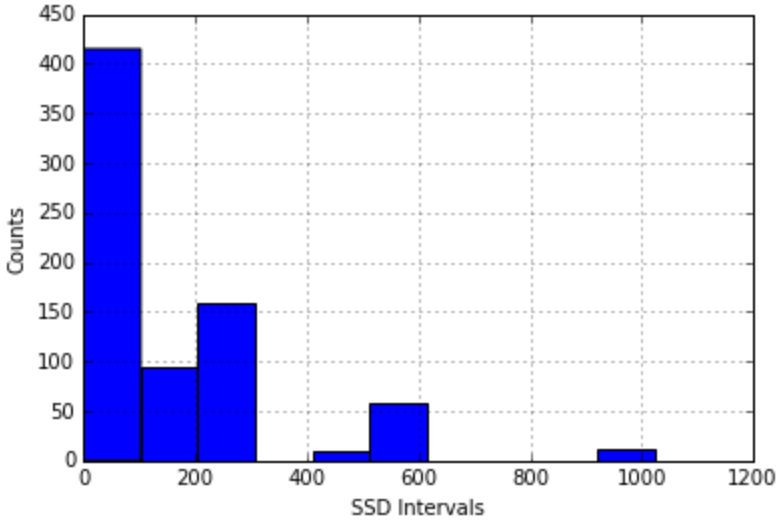
The variance of Pro Speed 0.22
The mean of Pro Speed 2.35

Ratings:



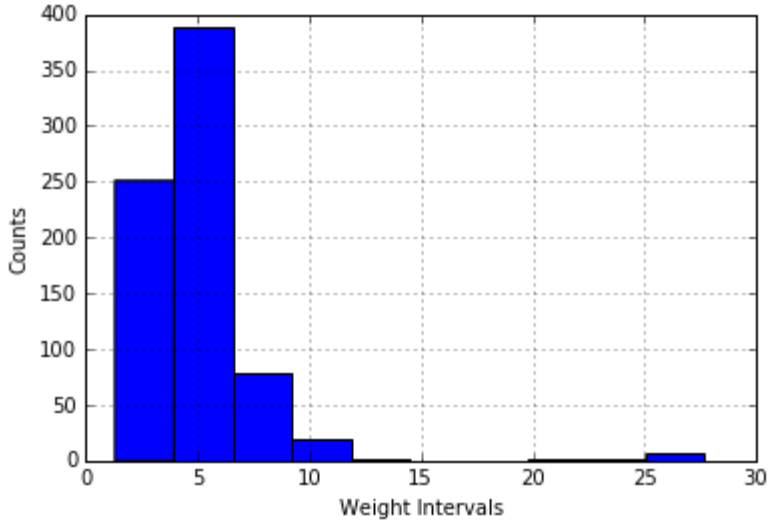
The variance of Ratings 4.58
The mean of Ratings 2.05

SSD:



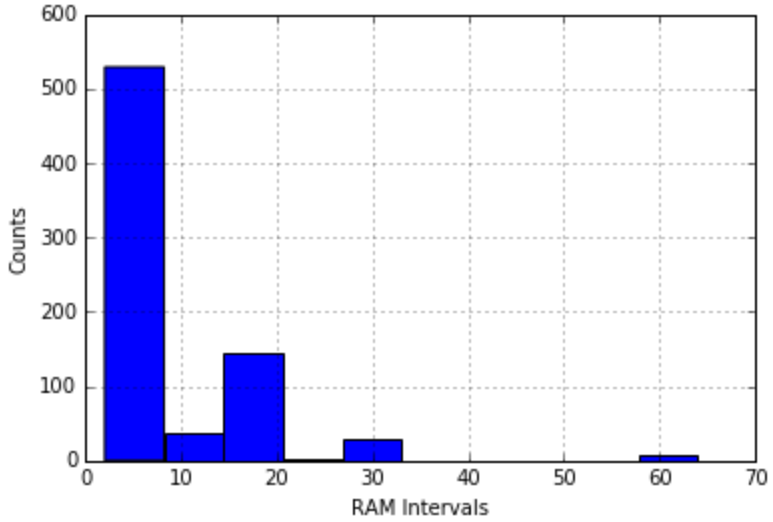
The variance of SSD 37585.37
The mean of SSD 135.09

Weight:



The variance of Weight 8.82
The mean of Weight 5.08

RAM:



The variance of RAM 72.29
The mean of RAM 9.82

Evaluation

Performance Measure

The performance measures are as below :

1) The estimator's score method : This was used to check the overall scoring of the estimator.
2) The metric functions R2 score, Negative Mean Absolute error, Negative Mean Squared error, Negative Median Absolute Error were used to check what is the prediction score and error of the model used for predicting.

Classifiers and Discussion

The classifiers and their parameter settings that were used are explained and discussed below:

1) DummyRegressor(strategy='mean') : This was used as a baseline model on which the rest of the classifiers were compared to.
2) LinearRegression(fit_intercept='True',Normalize='False')

   It is one of the most widely known and the very first modeling technique that comes up in Regression. There is a linear relationship between the features and target class for eg. SSD and Retail Price, RAM and Retail Price.

   There are however some evident problems seen that could cause errors in the prediction power of this model - the data is multicollinear(related to each other in one way or the other), therefore chances that the variance in the coefficients could be unstable even with minor changes. Linear Regression uses Least Squares Method.
3) Lasso Regression

   Although similar to Ridge regression, it also adds a penalty alpha to the size of the coefficients. It used L1 regularization method and shrinks coefficients to zero which helps in feature selection. It also helps in reducing the number of features when there is a correlation between them.

   The different alpha values we tried along with the coefficients for highest coefficient(OEM=Microsoft) and lowest coefficient(Screen Resolution=2736 x 1824) are as below:

   alpha = 0.1,1020.4,-2121.24

   alpha = 0.5,886.26,-915.92

   alpha = 1.0,778.00,-486.39

   alpha = 2.0,574.04,-0.0

   alpha = 5.0,384.7,-0.0

   alpha = 10.0,46.27,-0.0

One interesting find :

with alpha = 5.0

Graphics=NVIDIA : **290.87**

OEM=Microsoft : 384.70

with alpha = 10.0

Graphics=NVIDIA : **316.78 (becomes higher than OEM=Microsoft)**

OEM=Microsoft : 34.22

There is a change in the coefficient with the highest value. Increasing the value of alpha also decreases the model score. Out of the 64 input features, when using Lasso(alpha=10.0), only 20 features have non zero values.

4) Ridge Regression

We decided to use this model because of the multicollinearity of the input features. Errors are reduced here because of a regression bias or alpha that we add to it i.e it adds a penalty to the size of the coefficients. The larger the value of alpha, greater the amount of shrinkage and the coefficients becomes more robust.

The different alpha values we tried along with the coefficients for highest coefficient(OEM=Microsoft) and lowest coefficient(Screen Resolution=2736 x 1824) are as below:

alpha = 0.1,786.84,-1216.09

alpha = 0.5,589.72,-466.98

alpha = 1.0,480.4,-260.66

alpha = 2.0,368.58,-129.11

alpha = 5.0,234.04,-43.63

alpha = 10.0,152.5,-18.62

5) Elastic Net Regression

This is a hybrid of Lasso and Ridge and uses L1 and L2 regularization. It was used to experiment on the shrinkage properties it has on the coefficients. Using this model, the coefficients for highest coefficient(OEM=MSI) and lowest coefficient(Screen Resolution=1366 x 768 (HD)) which are different from the other models mentioned above.

Model score also decreases as the value of alpha increases.

6) Orthogonal Matching Pursuit

This model gave the lowest model score among all the others. Only 6 out of the 64 features had non zero values which was very interesting.

Evaluation Strategy

Because the model is to be ideally used for predicting the price of a laptop based on the given specifications, a train test split was done. Test_size was kept as 0.7.

Performance Results

MODEL	PARAMETERS	PERFORMANCE
Baseline : DummyRegressor	Strategy='mean'	Score : -6.9
LinearRegression	fit_intercept=True Normalize = False	Score : 0.79
Lasso	Alpha =1.0 Alpha =0.1 Alpha = 2.0	Score : 0.82 Score : 0.80 Score : 0.81
Ridge	Alpha = 1.0 Alpha = 0.1 Alpha = 2.0	Score : 0.82 Score : 0.81 Score : 0.82
Elastic Net Regression	Alpha = 1.0 Alpha = 0.1 Alpha = 2.0	Score : 0.76 Score : 0.81 Score : 0.74
Orthogonal Matching Pursuit	none	Score : 0.79

Cross Validation Results

MODEL	CV PARAMETERS	MEAN SCORES
LinearRegression	cv=10 cv=10, scoring='r2' cv=10, scoring='neg_mean_absolute_error' cv=10, scoring='neg_mean_squared_error' cv=10, scoring='neg_median_absolute_error'	0.76 0.76 -211.95 -114183.18 -134.74
Lasso	cv=10 cv=10, scoring='r2' cv=10, scoring='neg_mean_absolute_error' cv=10, scoring='neg_mean_squared_error' cv=10, scoring='neg_median_absolute_error'	0.81 0.81 -209.28 -102662.42 -139.65
Ridge	cv=10 cv=10, scoring='r2' cv=10, scoring='neg_mean_absolute_error' cv=10, scoring='neg_mean_squared_error' cv=10, scoring='neg_median_absolute_error'	0.81 0.81 -208.6 -102380.24 -136.64
Elastic Net	cv=10 cv=10, scoring='r2' cv=10, scoring='neg_mean_absolute_error' cv=10, scoring='neg_mean_squared_error' cv=10, scoring='neg_median_absolute_error'	0.80 0.80 -212.84 -110348.33 -145.80
Orthogonal Matching Pursuit	cv=10 cv=10, scoring='r2' cv=10, scoring='neg_mean_absolute_error' cv=10, scoring='neg_mean_squared_error' cv=10, scoring='neg_median_absolute_error'	0.74 0.74 -252.05 -135455.27 -172.6

Top Features

Importance was measured based on the value of the coefficients of the model.

Top feature : OEM = Microsoft

Least important feature : Screen Resolution = 2736 x 1824

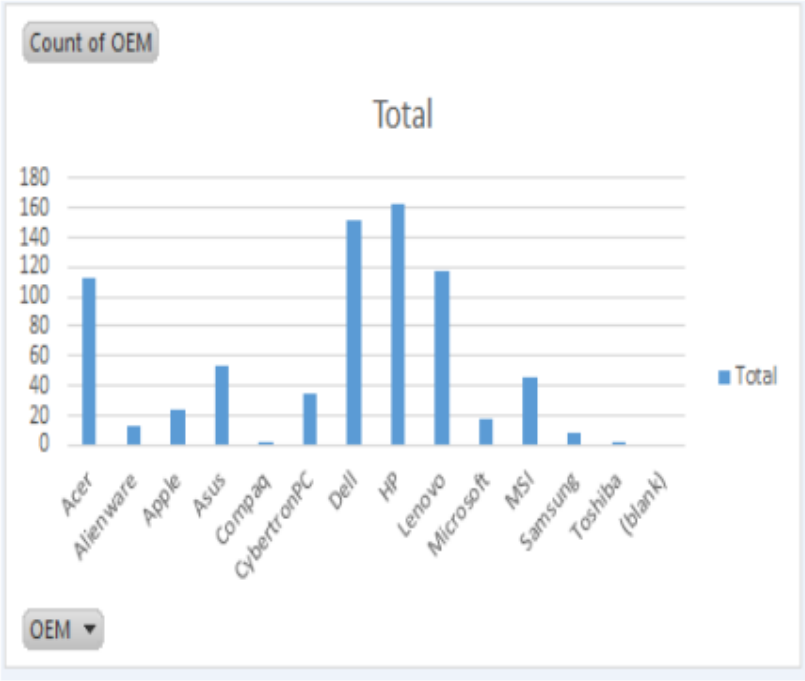
Interesting/Unexpected Results

Laptops with the OEM as Microsoft are only 17 instances in the data set but OEM being Microsoft was a top feature. Goes to show that Microsoft has devices with prices that can reflect the general trend in the overall pricing strategy for the entire market. But when the model predicts the prices for Microsoft it performs very poorly.

Model Score of OEM=Microsoft : -0.53 (!!!)

Model Score of OEM = Dell : 0.72(second highest number of instances)

Model Score of OEM = HP : 0.76 (highest number of instances)



Row Labels	Count of OEM
Acer	112
Alienware	13
Apple	24
Asus	53
Compaq	1
CybertronPC	35
Dell	152
HP	163
Lenovo	117
Microsoft	17
MSI	46
Samsung	8
Toshiba	1

The project was initially done taking only 16 features (to consider the small size of data). Accuracy was 70% with most weighted feature being 'OS=Mac OS'. When the number of features was increased to 65, accuracy increased to only 79% with the most weighted feature being 'OEM=Microsoft'.

Contributions of Each Group Member

The contribution from each of the members was equal in terms of technical work and analytical effort. More details are given below:

Chris Dani

- Performed web scraping of the laptop lists from Bestbuy website.
- Analyzed the model's performance with the data and coding.
- Worked in finding interesting cases.

Mabin Mariam Joseph

- Did an initial analysis of the data and worked on manual cleaning to avoid losing the retrieved data.
- Chose the models to be used for the task and performed coding.
- Worked in finding interesting cases.

Overall both the members were involved in the success of the project at each phase.

Conclusion

The most weighted feature is the OEM being Microsoft, which goes to show that Microsoft has the most expensive laptops. OEM being Toshiba has the least coefficient value showing that Toshiba makes the cheapest laptops for the given values.

For a linear model using only those data in the OEM category and then test data from the same OEM , train (where n > 5):

OEM	Score	#N
OEM=Acer	0.799553	112
OEM=Alienware	-0.5063	13
OEM=Apple	0.700215	24
OEM=Asus	0.483615	53
OEM=Cy-bertronPC	0.659578	35
OEM=Dell	0.156065	152
OEM=HP	0.766429	163
OEM=Lenovo	0.285189	117

OEM=MSI	0.166704	46
OEM=Microsoft	-11.5633	17
OEM=Samsung	-11.4447	8

Prediction when the model is built with all the data and each OEM is tested

OEM	**Score**	Beta value
OEM=Acer	0.924952	-191.706
OEM=Alienware	0.782343	-181.572
OEM=Apple	0.76608	114.8361
OEM=Asus	0.928605	-117.49
OEM=Compaq	0	-6.82E-13
OEM=CybertronPC	0.468084	261.0228
OEM=Dell	0.759187	-239.705
OEM=HP	0.701094	-205.938
OEM=Lenovo	0.740266	-127.45
OEM=MSI	0.485378	368.6167
OEM=Microsoft	-1.66893	927.5689
OEM=Samsung	-0.38379	-242.5
OEM=Toshiba	0	-537.916

Sample prediction for few values in the best and worst prediction :

Acer

Retail Price	OS	Pro Speed	RAM	Screen Size	Touch Screen
389.99	Windows 10	2.3	4	15.6	No
699.99	Windows 10	2.3	8	14	No
749.99	Windows 8.1	1.7	8	15.6	Yes
279	Chrome OS	1.6	4	11.6	Yes
399.99	Windows 10	2.3	4	15.6	No

Weight	Ratings	OEM	HDD	SSD	Convertible	Refurbished	Pro Brand
5.29	4.3	Acer	1000	0	No	No	Intel
4.19	5	Acer	0	256	Yes	No	Intel
5.29	5	Acer	1000	0	Yes	No	Intel
2.76	4.4	Acer	0	16	Yes	No	Intel
5.29	0	Acer	500	0	No	No	Intel

Screen Resolution	Graphics	Predicted	Actual	#	% accuracy
1366 x 768 (HD)	Intel	508.01512306	389.99	158	0.7676739969
1920 x 1080 (Full HD)	Intel	864.66740362	699.99	150	0.8095482692
1920 x 1080 (Full HD)	Intel	916.15292728	749.99	237	0.8186297044
1366 x 768 (HD)	Intel	317.2902793	279	32	0.8793209821
1366 x 768 (HD)	Intel	545.34611554	399.99	466	0.7334608033

MSFT

Retail Price	OS	Pro Speed	RAM	Screen Size	Touch Screen
599.99	Windows 10	1.6	4	10.8	Yes
1599.99	Windows 10	2.3	8	12.3	Yes
1899.99	Windows 10	2.4	8	13.5	No
999.99	Windows 10	2.4	4	12.3	Yes
999.99	Windows 10	1.1	4	12.3	Yes

Weight	Ratings	OEM	HDD	SSD	Convertible	Refurbished	Pro Brand
21.92	4.5	Microsoft	0	128	Yes	No	Intel
27.7	4.4	Microsoft	0	256	Yes	No	Intel
3.34	4.4	Microsoft	0	256	Yes	No	Intel
27.5	4.5	Microsoft	0	128	Yes	No	Intel
27.04	0	Microsoft	0	128	Yes	No	Intel

Screen Resolution	Graphics	Predicted	Actual	#	% accuracy
1920 x 1280	Intel	3087.84048873	599.99	479	0.1943073168
2736 x 1824	Intel	1532.57928047	1599.99	78	1.0439851435
0	NVIDIA	2059.68239356	1899.99	84	0.9224674668
2736 x 1824	Intel	1309.96963592	999.99	20	0.7633688389
2736 x 1824	Intel	1222.27985254	999.99	116	0.8181350596

References

The below web sites were referred to :

1) http://scikit-learn.org/stable/modules/linear_model.html
2) https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/