MOBILE PRICE PREDICTION

Multiclass classification problem

TABLE OF CONTENTS

Submitted by: Sushanth S (MAIB-2021)

INTRODUCTION:

XYZ company manufactures various mobile phones customized to meet a customer segment. Company has invested heavily on artificial intelligence. Now, XYZ company wants it price tagging and mobile classification to be automated in the assembly line.

XYZ has provided the AI scientists historical data of mobile price categories based on the custom-built features.

Mobile price dataset consists of multiple attributes such as battery power, clock speed, RAM, dual sim, inbuilt memory etc. Based on the input variable the price of mobile needs to be predicted.

DATASET:

Below is a sample of dataset provided by the XYZ company

battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	...	px_height	px_width
842	0	2.2	0	1	0	7	0.6	188	2	...	20	756
1021	1	0.5	1	0	1	53	0.7	136	3	...	905	1988
563	1	0.5	1	2	1	41	0.9	145	5	...	1263	1716
615	1	2.5	0	0	0	10	0.8	131	6	...	1216	1786
1821	1	1.2	0	13	1	44	0.6	141	2	...	1208	1212

ram	sc_h	sc_w	talk_time	three_g	touch_screen	wifi	price_range
2549	9	7	19	0	0	1	1
2631	17	3	7	1	1	0	2
2603	11	2	9	1	1	0	2
2769	16	8	11	1	0	0	2
1411	8	2	15	1	1	0	1

The highlighted column is the feature to be predicted for different models.

Auditing the dataset, we get to know:

1. Most of the feature are numerical variable.

2. Some features such as blue, dual sim, 4G, Number of cores, 3G, Touch screen and Wi-Fi are categorical values.
3. Price data contains 4 different categories.
4. Data checked for null values.
5. After completing the feature engineering step on data, the final data was split into train and test data using a stratified slip approach.

MODEL DEVELOPMENT:

The following classification models were tested for the above dataset

1. Decision Tree Classifier
2. Random forest
3. Gradient booster
4. Gaussian NB
5. K Nearest Neighbour

HYPOTHESIS:

As the correct prediction of mobile price category is important for the business.

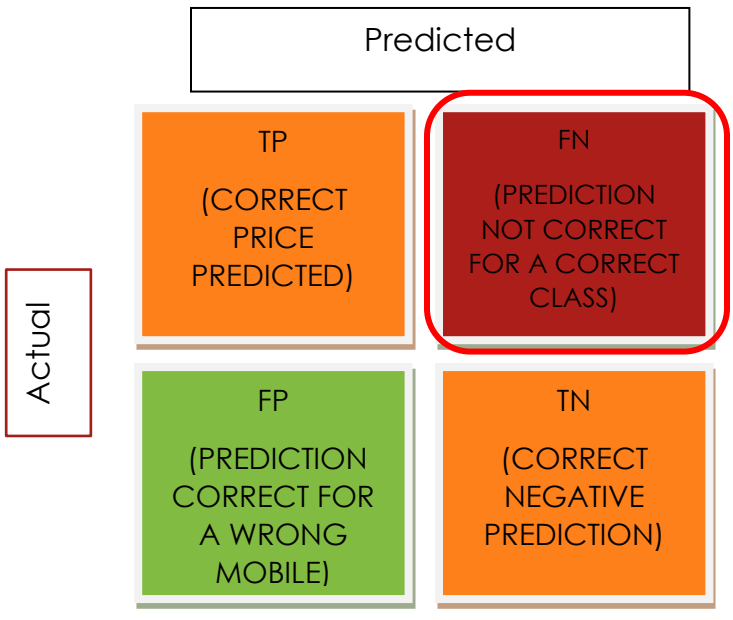
If the prediction price is not correct (Negative) and the prediction is false, then it can demean the business objective

So, in our case False negative is a very important measure. The lower the false negative the better the outcome of the model.

Alternate Hypothesis: Select the model with highest Recall and Best fit

Null Hypothesis: Reject the model when the model is not a best fit and recall value is lower.

However, **for mobile price prediction the accuracy and F1 score is more important than FN.**



MODEL EVALUATION

Comparison of performance measures:

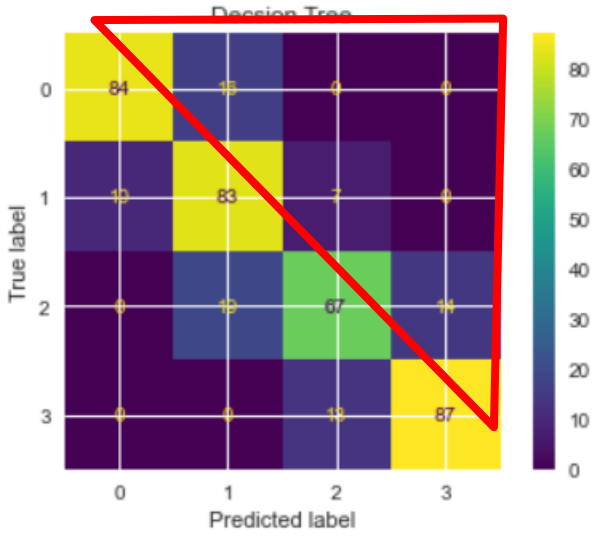
	Decision Tree Classifier	Random Forest	Gradient Booster technique	Gaussian NB	KNN
Accuracy Training	86.1	100	**96.8**	81.4	94.9
Accuracy Testing	80.2	86.5	92.7	83.3	91.5
precision	80.7	86.9	92.8	83.3	91.5
recall	80.25	86.5	**92.7**	83.25	91.5
F1 Score	80.24	86.5	92.7	83.29	91.5
AUC	95.07	97.3	**99.17**	96.3	98.5

From above table it is evident that **Gradient Booster technique is the best** performing model.

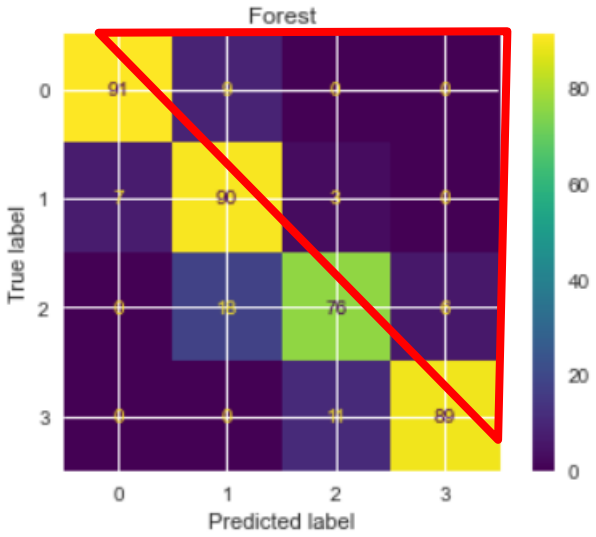
All other models can be rejected based on the NULL hypothesis.

Confusion Matrix of each model: (Area of focus is highlighted)

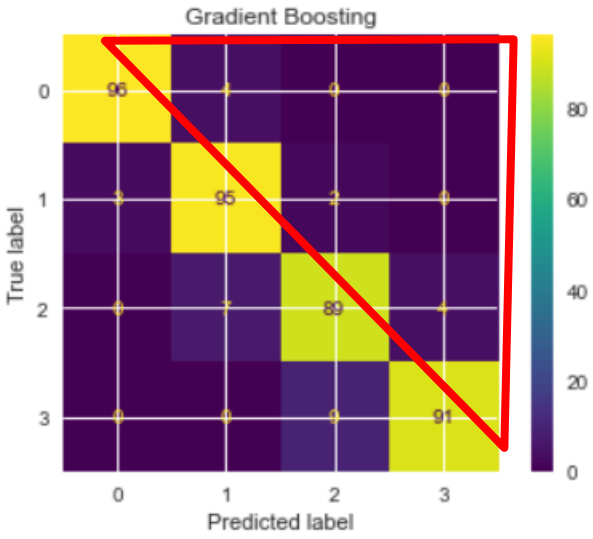
1. Decision Tree Classifier



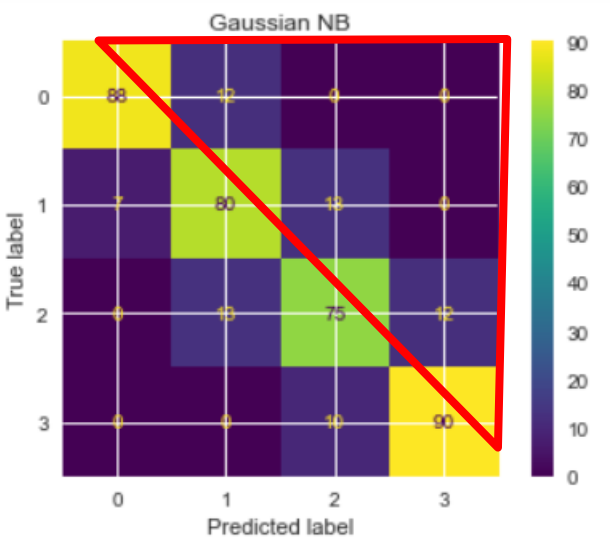
2. Random Forest



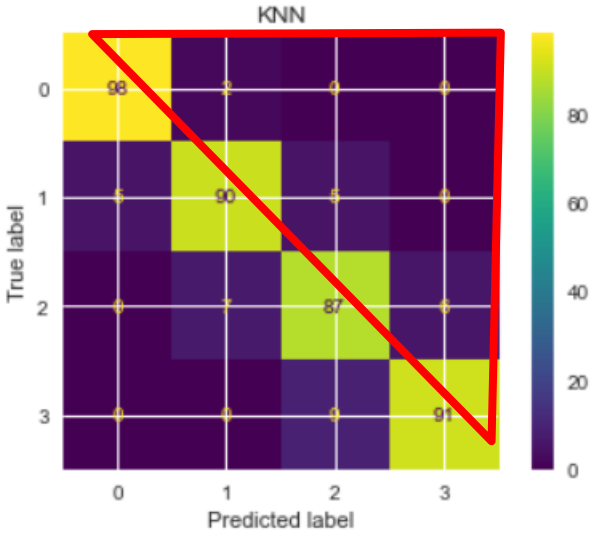
3. Gradient booster



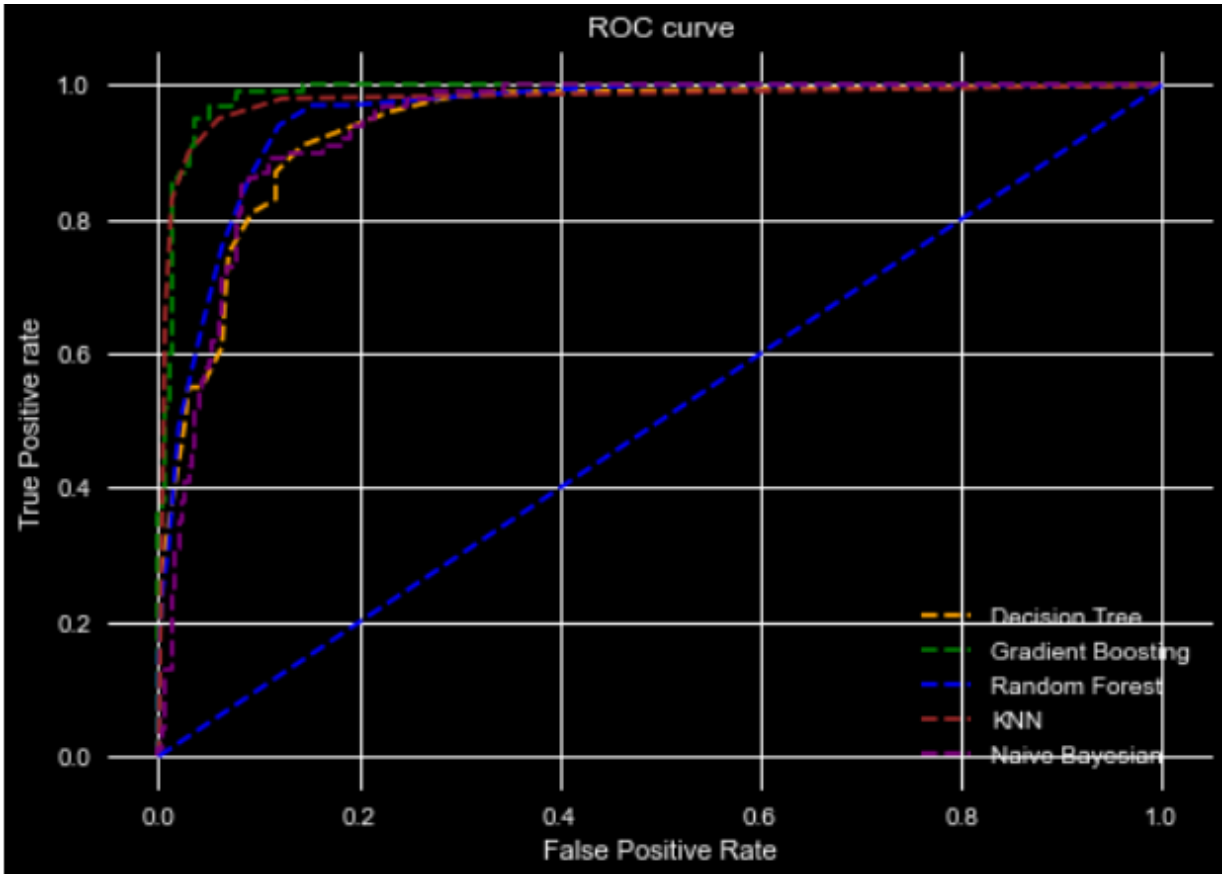
4. Gaussian NB



5. K Nearest Neighbour



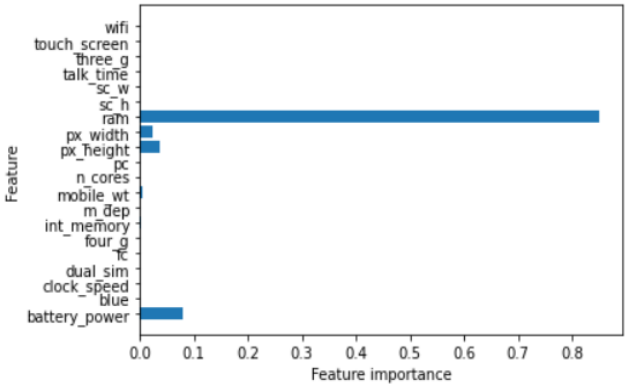
ROC Curve Comparison:



CONCLUSION

The data set consists of a mix of many numerical as well as categorical values. The best performing model is **Gradient Boosting Technique**.

Gradient boosting gives a best fit with Training accuracy – 96.8 % and testing accuracy 92.7 %. Only a 4.1 % difference between the accuracy gives a good confidence level.



Gradient boosting algorithm refined the features and developed the model based on Ram, pixel width, pixel height and battery power as the important feature.

Further when Recall, the parameter of importance, is check we find **92.7%** highest among all the models.

The ROC plot also clearly indicates **Gradient boosting** as the model with highest area under its hood. **AUC score – 99.17%**