Smartphone Pricing Range Classification

Team *Apex Legends*
Bowen Wang(bw2716), Zhengji Wang(zw2785), Serena Ding(sd3517),
Tianyi Li(tl3094), Nick Chen()

Abstract

With the development of technology and an increasingly competitive mobile market, new mobile sellers are finding it hard to find acceptable prices for phones with certain combinations of features. In this project, by finding the relationship between functionality and prices, we explore and classify the phone prices using Logistic Regression, KNN and Random Forest models

Understanding the raw dataset & Features

As we've found that there exist zero values in height and pixels as an example, we should get rid of these data as they are possibly invalid data. These data are important for determining the price since they are associated with fixed cost of producing phones and hence we could definitely not add "reasonable data" artificially. Besides, "blue", "dual_sim", "four_g", "three_g", "touch_screen", "wif" are all binary variables and hence one-hot encoding is not necessary. After removing these invalid zero values, we fully prepared our raw dataset for further analysis.

Before starting features analysis, we check if there is any missing or duplicating value in any part of the data set, since they could be disturbing and could possibly affect the accuracy of the prediction of our model. And, we want to have a general sense about the dataset and relationship between features and price range.

Numerical Features

To begin with, we use the seaborn package to demonstrate the correlation between all numerical features and the price range using boxplot. It seems that "battery_power", "pc", "px_height", "px_width", "ram" is positively correlated to prices since as the value of these features increases, prices of phones tend to be higher. These correlations are reasonable as battery, primary camera mega pixels and pixel resolution are features associated with costly parts of a phone, while other features do not demonstrate an observable correlation with prices.

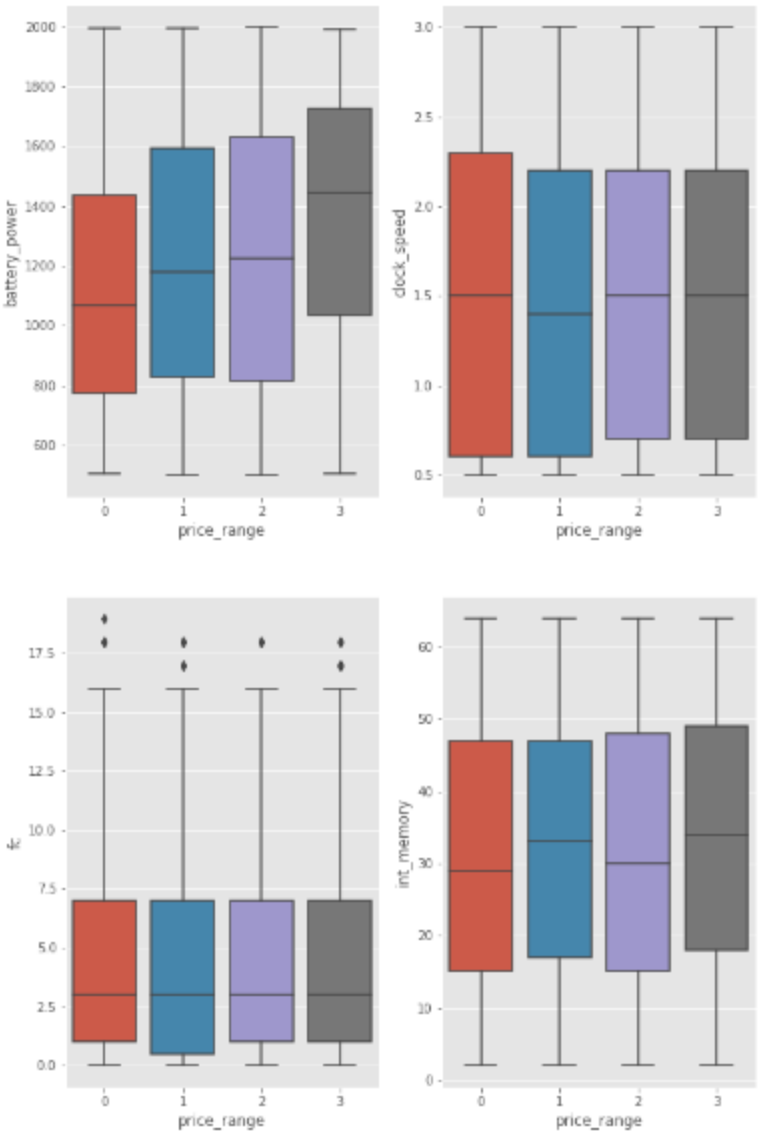


Fig. 1 Boxplot of Numerical Features (Partial)

We further present the heatmap of correlation between these features to find if highly correlated features exist using heatmap as a way of demonstration. Apparently, “pc” and “fc”; “px_height” and “px_width” are 2 groups that are correlated to each other respectively. These correlations are expectable since camera pixels and resolution influence each group of these features in a similar way.

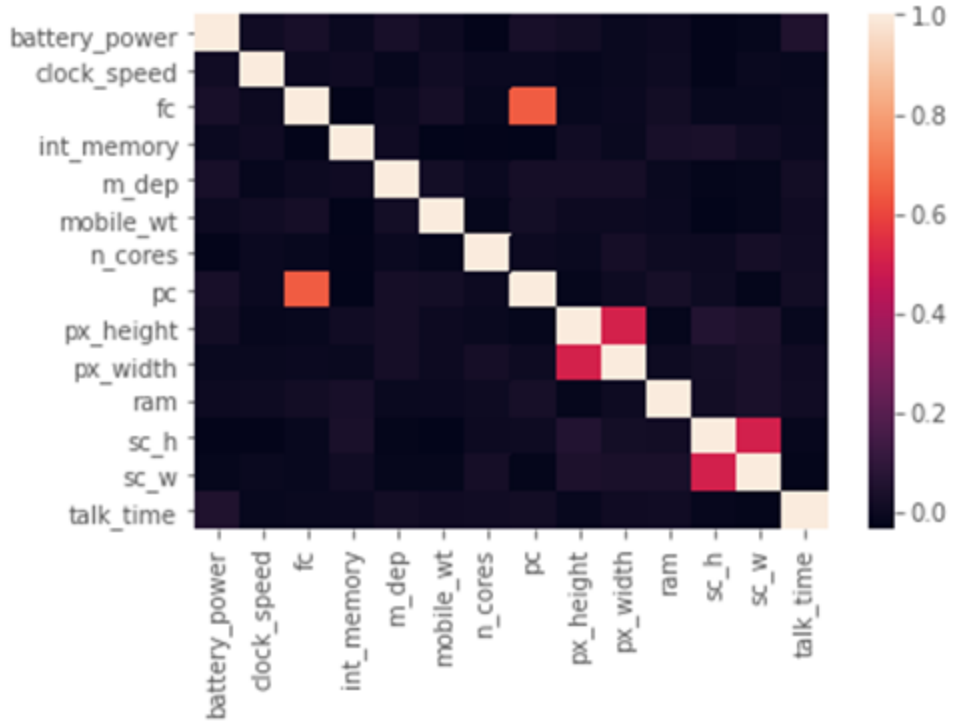


Fig.2 Heatmap of Numerical Features

Categorical Features

Then, we would want to understand the categorical features. For this type of features, instead of using boxplot, we used count plot to more clearly demonstrate the functionality distribution in each of the price range sectors.

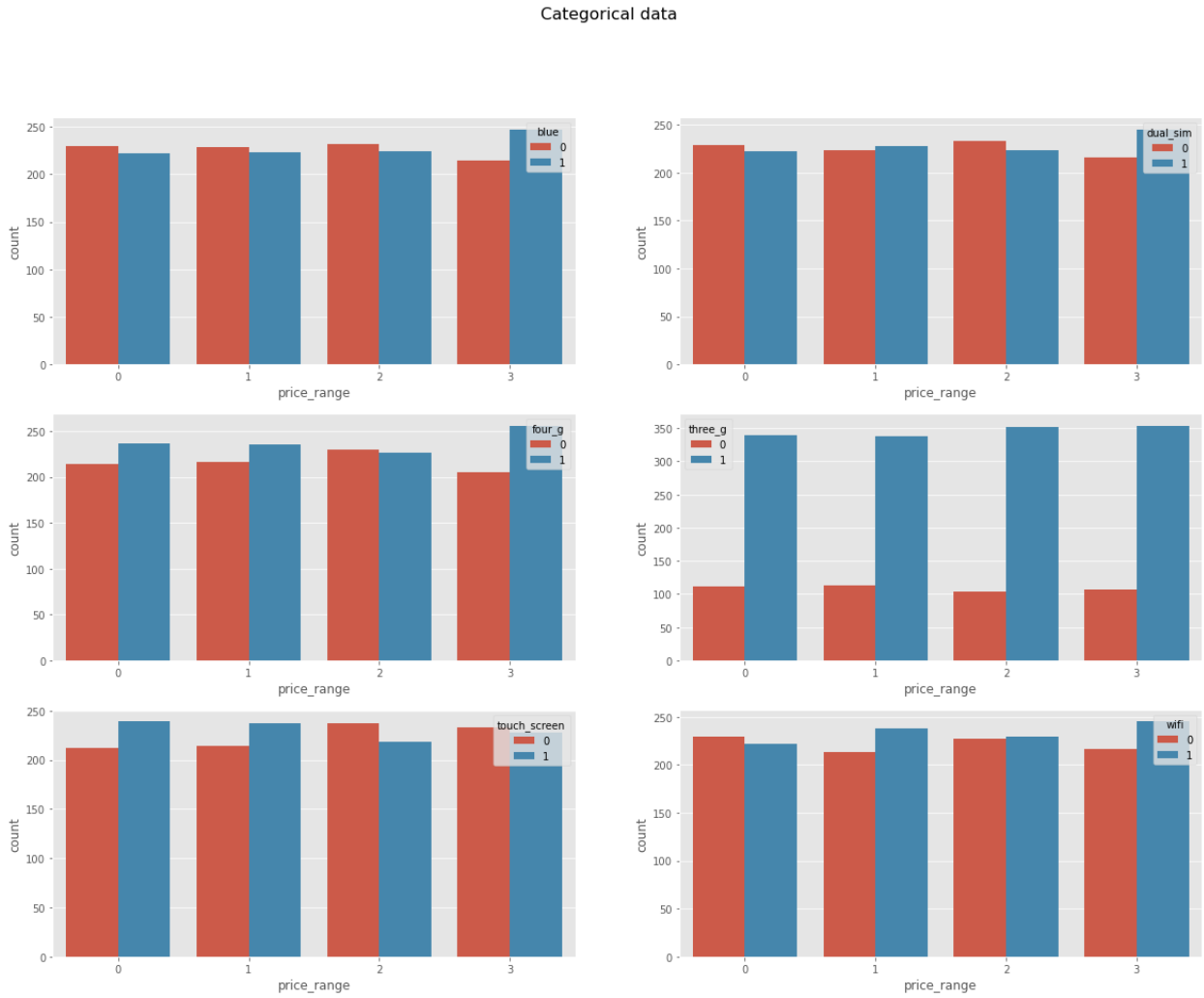


Fig.3 Categorical Features

As the presence of “blue”, “dual_sim”, “four_g” and “wifi” increases, the target column price range that the phone belongs to goes from 0 to 3.

Feature Engineering

First, apply Logistic Regression, KNN and Random forest on the raw trained dataset and look at its k-fold cross validation MAE score for each fold. This would be a control group for the feature selection criterion. After selecting and combining features, what needs to be achieved is an improved or same MAE score as model performance is still vital.

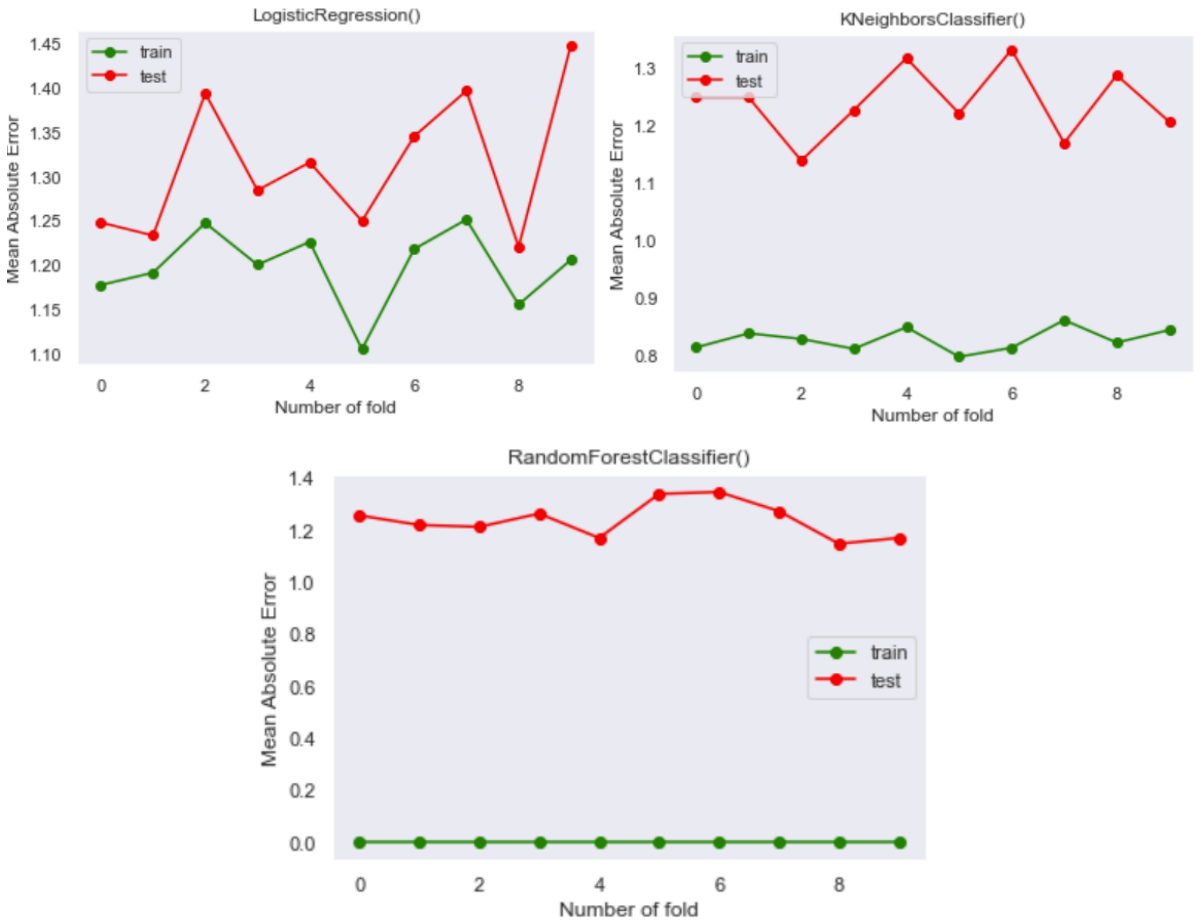
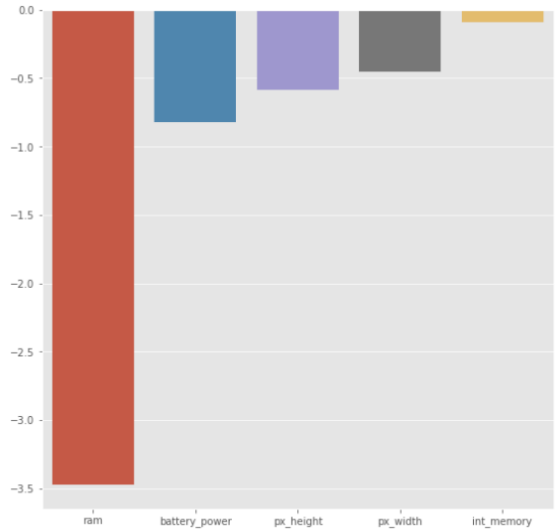


Fig.4 Initial MAE scores plots

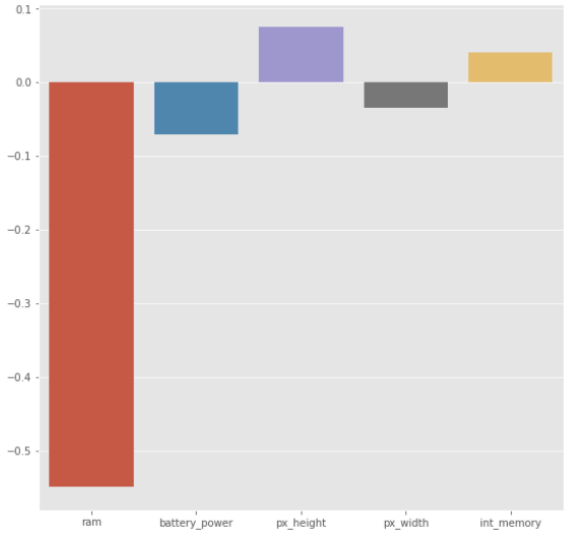
After data exploration and preprocessing,lasso regression, ridge regression and random forest feature importance are used to shed light on which features need to be processed and which are vital and should be untouched.

After temporarily standardizing the data and applying lasso regression on the raw trained dataset, lasso regression assigns the 0 coefficient to fc , pc , blue, n_cores in different price ranges. And the 5 most important features for each price groups are below:

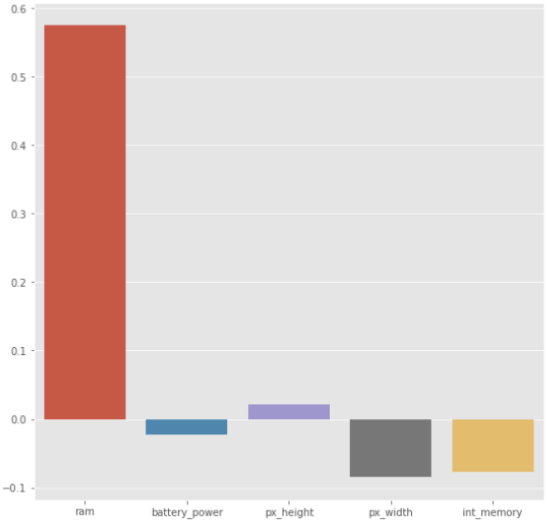
Price range: 0



Price range: 1



Price range:2



Price range: 3

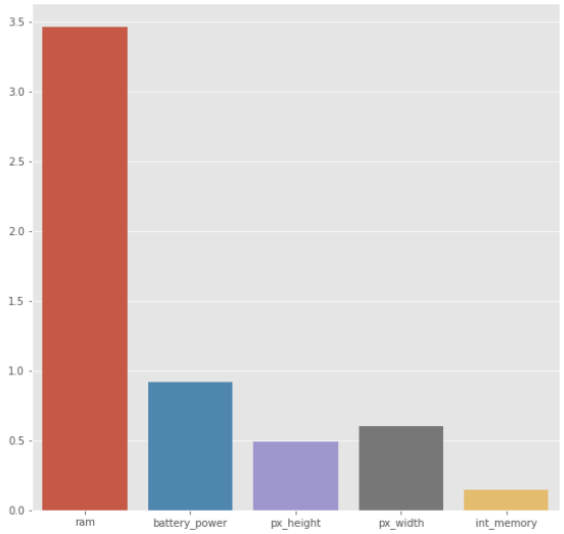


Fig.5 Logistic regression feature importance plots

So it is better not to touch ram, battery power, pax_height, px_width and int_memory.

Ridge regression wouldn't be so aggressive and assign coefficients 0 easily, but it would also shed some light on which features are important and which features are unimportant based on coefficients. Basically it gave the same result as lasso regression.

Finally looking at random forest, the feature importance it gave is as follows:

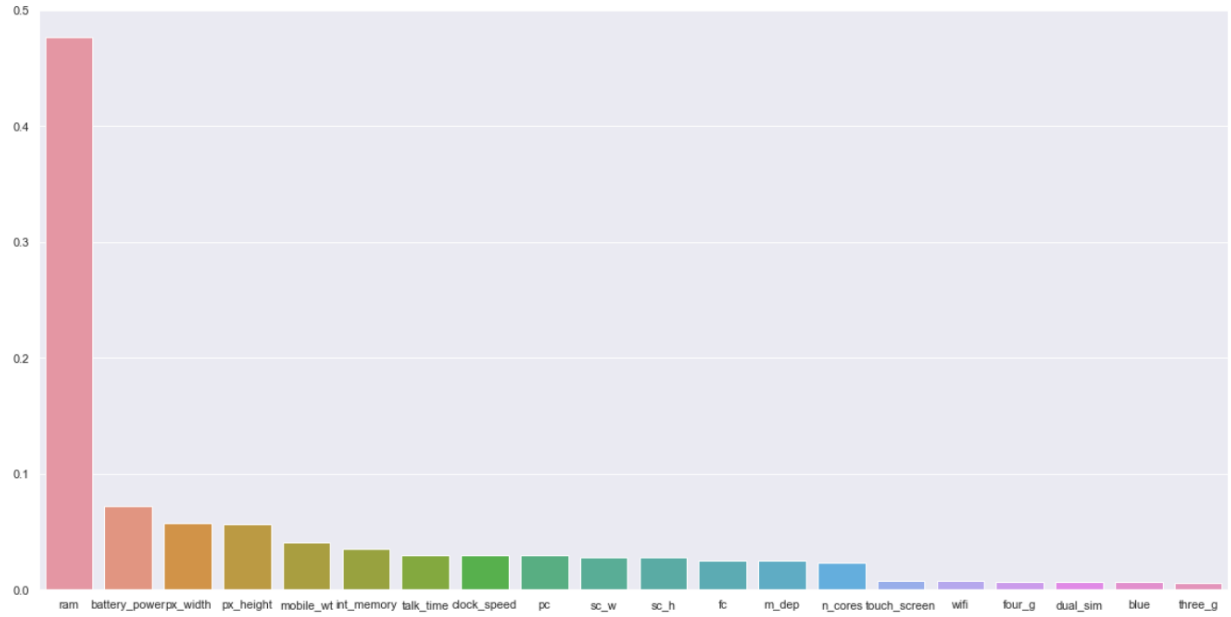


Fig.6 Random Forest Features

It is obvious that touch_screen, wifi, four_g,dual_sim, blue, three_g are all considered as unimportant.

After experimenting with different features dropping choice with the information given above, The decision is made as follows:

1. For logistic regression, sc_h and sc_w are combined into sc_area, because it makes sense to do so, as customers are more likely to care about the overall size of the phone screen instead of just screen width and screen height. Also pc, fc(highly correlated and low importance), dual_sim, blue, wifi, three_g, four_g, touch_screen, talk_time as they all have low coefficient in ridge regression and some of them are 0 in lasso regression.
2. For KNN, after several experiments, the decision is made that it might be better not to drop any feature because doing so would reduce performance. So sc_h and sc_w are combined into sc_area , but no columns are dropped.
3. For random forest, three_g, four_g, blue, dual_sim, touch_screen, wifi, n_scores, fc are dropped.

After all these preprocessing, the result of the raw model on trained raw dataset look like these:

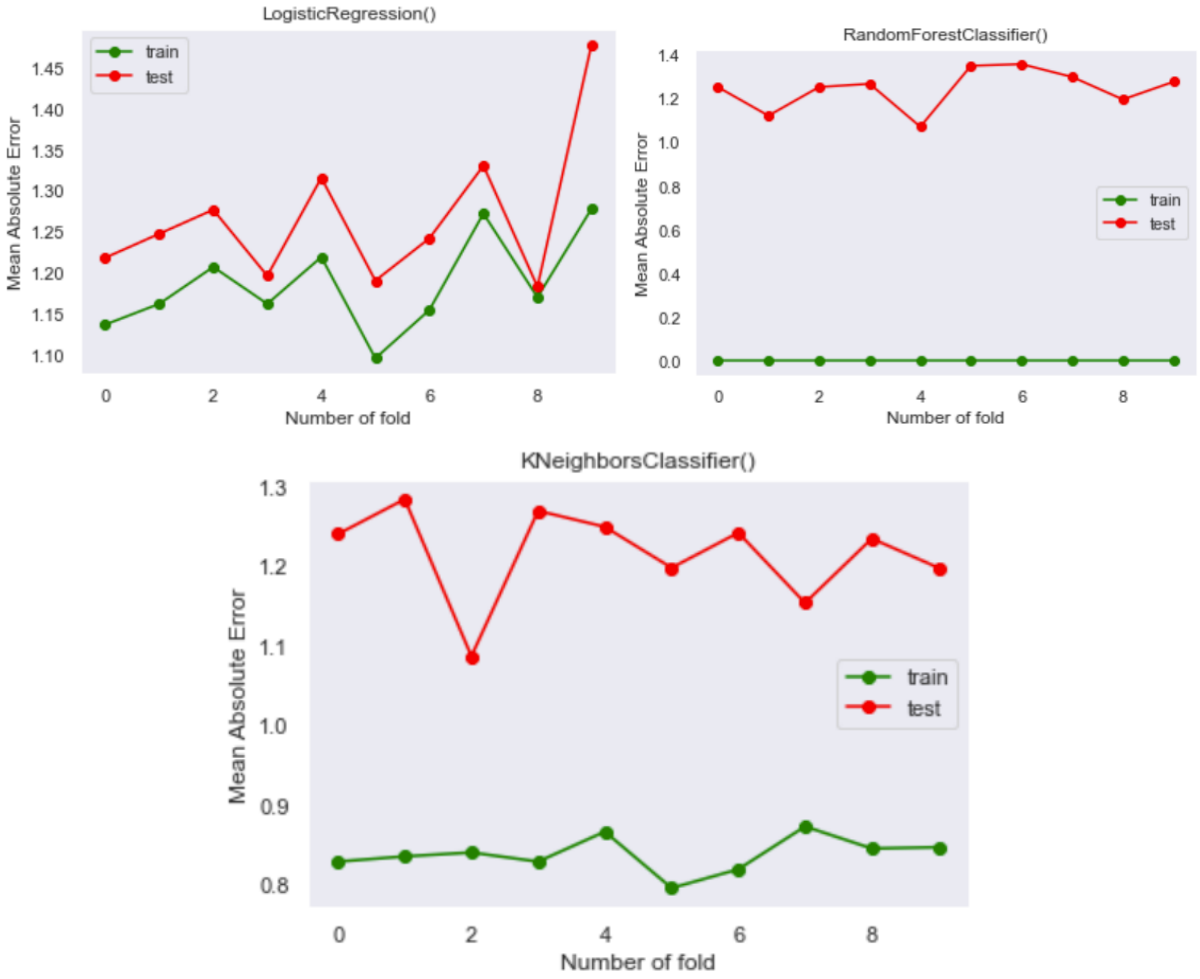


Fig.7 MAE scores after feature engineering

As is shown in the graph, the k-fold cross validation MAE score didn't change much, so this feature preprocessing can be counted as a success.

Standardized or not

We use k-fold cross validation average accuracy score to determine whether standardized or unstandardized data is suitable for each model.

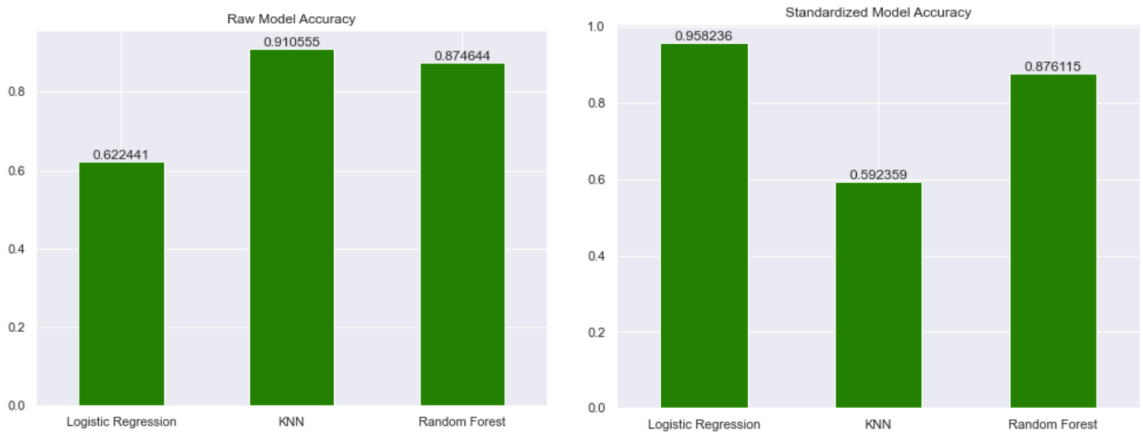


Fig.8 Model accuracy without/with standardization

It turns out logistic regression is most suitable for standardized data, KNN and Random Forest are better for unstandardized data.

Parameter Tuning & Model improvement:

We have successfully conducted feature importance analysis and data standardization. Then we did a Gridsearch algorithm to find the best parameters for each model and thus output the best models' performance. The best model scores are shown below:

Model Score	Logistic Regression	K-NN	Random Forest
Train	0.844619	0.928884	0.885650
Test	0.850549	0.927473	0.894505

Fig.9 Scores of Best Models

Regarding the best parameters, our Logistics regression chooses C = 1.0 as hyper parameters and Lasso L1 regularization as the penalty; random forests chooses 10 as maximum depth and 60 estimators; KNN 11 as its number of neighbors. We have visualizations for these process:

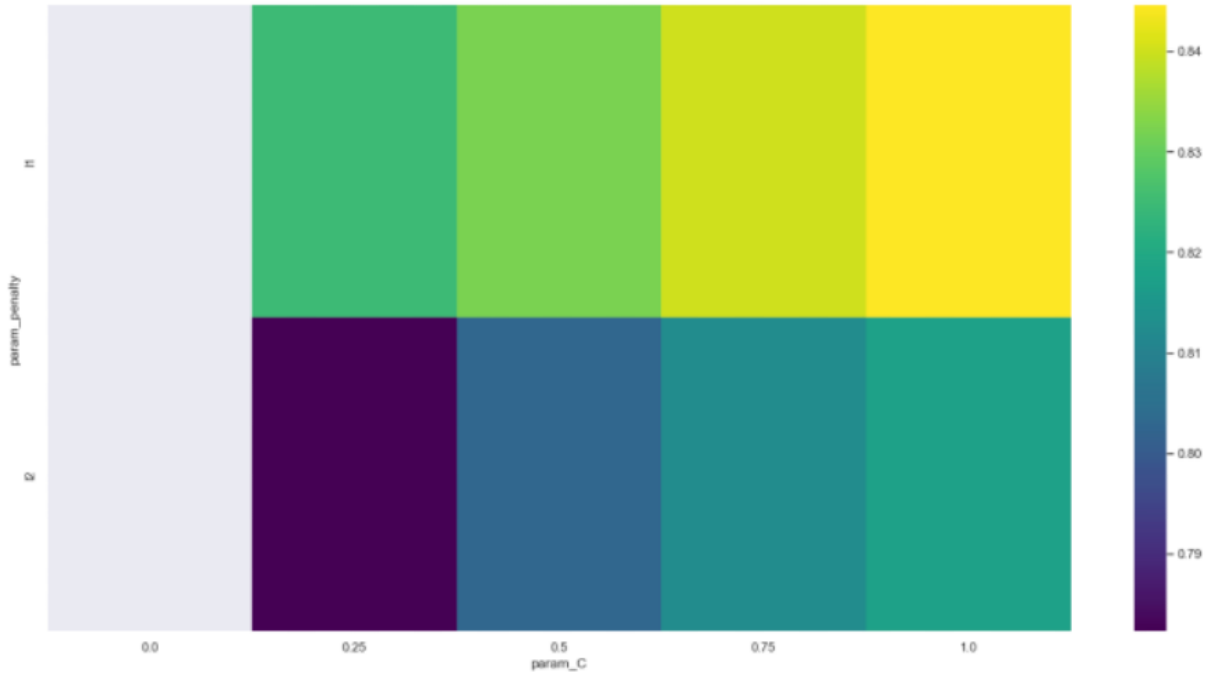


Fig.10 LR Heat Map

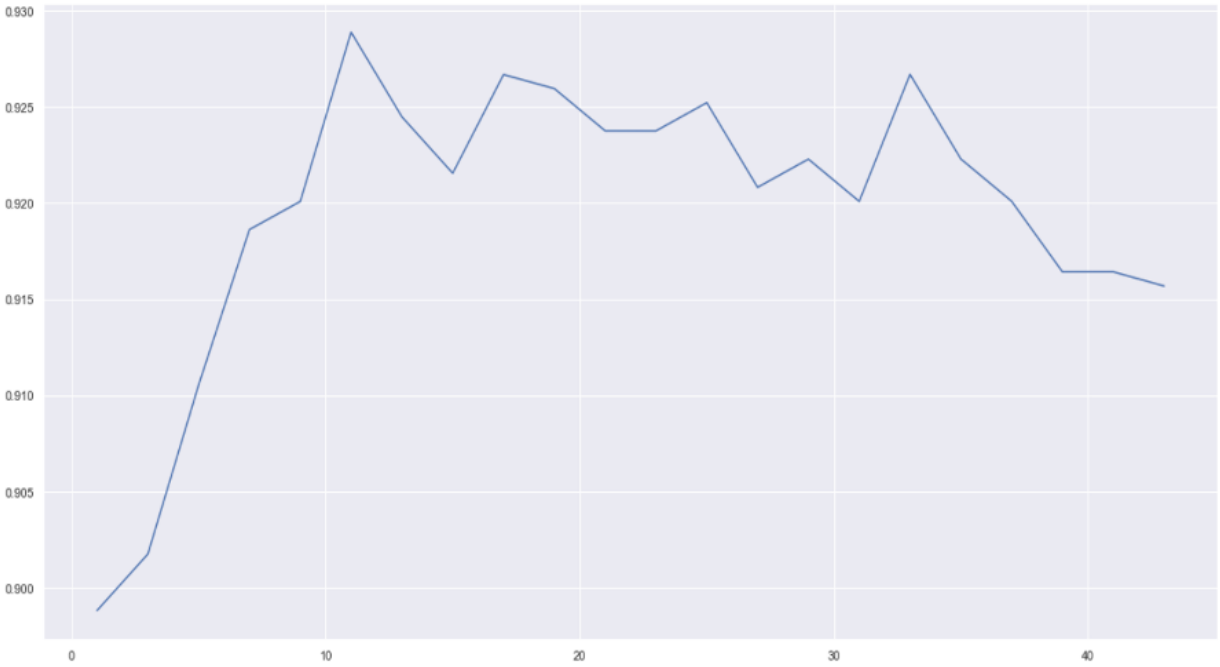


Fig.11 Optimal K

Then, we checked model accuracy again and find they all had improvements in different levels.The diagrams are shown below

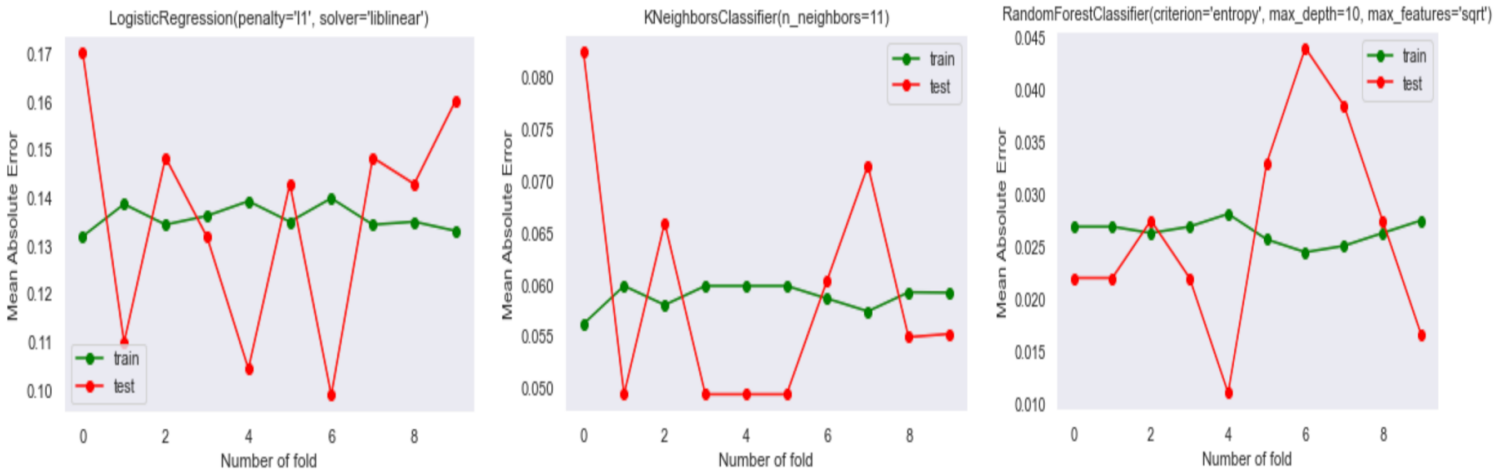


Fig.12 Check overfitting

Finally, we output confusion matrix as our metric to pick the best model to perform our task:

	Logistic Regression	K-NN	Random Forest
Accuracy	0.970149	0.955556	0.963801
Precision	1.000000	0.955357	0.980769
Recall	0.931818	0.955357	0.944444

Fig.13 Confusion matrix

Conclusions & Business Implications

Given our model performances, we finally choose the random forest model to be our target model in this project. To sum up our analysis, we believe that Ram, battery power and screen quality(both pixels and screen size) are determining factors when deciding which market a specific phone type that is being designed or experimentally manufactured should be placed in. Brands could then consider those data points carefully and construct a proper pricing strategy on launching new smartphone products. Thus, their brand agendas can be more easily achieved and their market shares can also be well maintained. So, we believe our project successfully delivers a valuable idea and would help a lot for smartphone manufacturers.