**Retail Price Classification**

Victoria Hall

July 24, 2022

**Retail Sales Forecasting**

The purpose of this project is to use a series of features about mobile phones to help determine what price range a phone should be in for a fledgling business. The data, which came from Kaggle.com (Sharma, 2018), was cleaned and prepped, then used to test two main classifier modeling techniques wherein numerous parameters were examined. The results of the best performing model should be used to help guide assigning the appropriate price range for a cell phone.

**Introduction**

**Business Problem**

A new fledging cell phone company is opening in a very crowded and competitive space. Companies like Apple and Samsung dominate the field but the market is huge with almost all Americans owning a smart phone so there are opportunities to grab market share and become a player in this industry (Pew Research Center, 2021). Staying competitive will be extremely important to the success of the company and a major aspect of that competitive edge is appropriately pricing product. To help identify what the price of a new phone should be, I will be using data on existing smart phones including their specs and current prices to recommend a price range for new phones in the business.

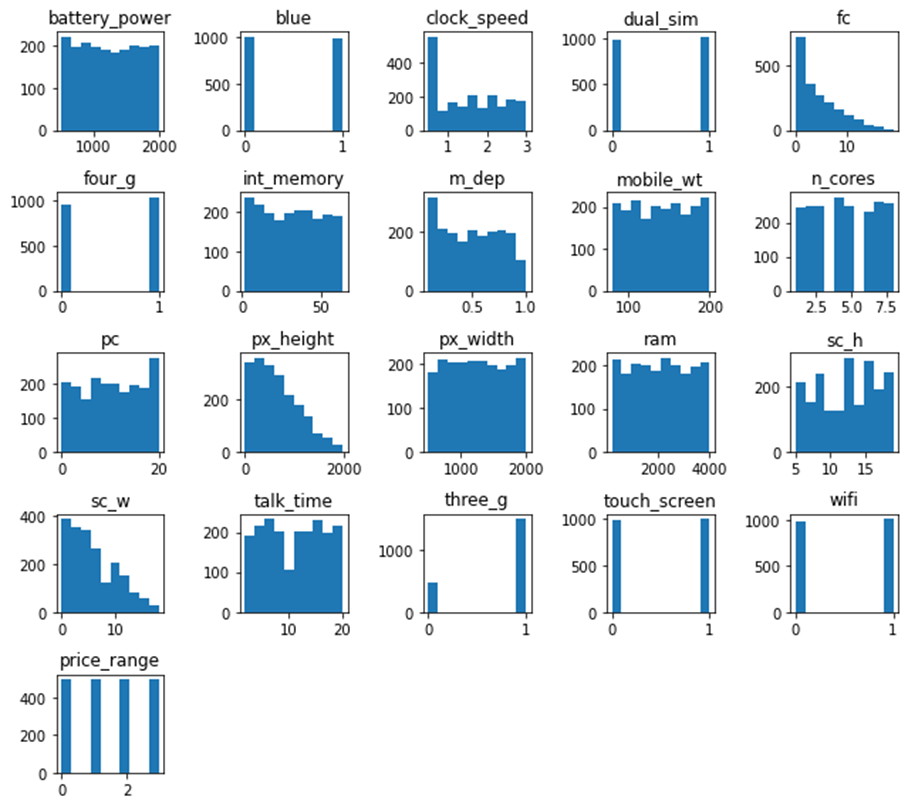
**History and Data**

The data for this project came from Kaggle.com as a classification task challenge. It was originally posted four years ago (Sharma, 2018). The data is contained in two flat files, one of which contains 2000 rows of data that is to be used to train the models. The other file contains 1000 rows for validation. The columns contain features about the smart phones include specs about the phone and the current price bucket. The specs of the phones, which will be used as features to teach the machine learning models include battery power, Bluetooth, speed, memory size, the size of the phone, and information about the camera. The pricing buckets are labeled as 0, 1, 2, and 3 and they stand for low cost, medium cost, high cost, and very high cost respectively.

In order to start training and testing models, there was some data preparation and exploration that had to be done. After inspecting the initial data frame, it was relatively clean. The only changes were that some columns were changed to categorical objects.

Bar chart

Description automatically generated with low confidenceThe exploration phase gave important insight into what types of models could be built and how they should be evaluated. First, there was an even split between the four price change classes.

There were also a number of features that evenly distributed. About half of the phones were 4G enabled, Bluetooth enabled, contained dual sim cards, had touch screens and Wi-Fi capabilities. There was little concern for outliers as well. 

The final check was for correlation among the features. The features with correlation between height and width for the screen and height and width of the pixels. Despite this, I do not have concerns about multicollinearity based on the modeling techniques that will be tested. Random Forest models use bootstrap and feature sampling; therefore, it is not affected by multicollinearity because it is picking different sets of features for different models and each model sees a different set of data points. This greatly reduces the risk for multicollinear features getting picked together (Raj, 2020). Similarly, XGBoost models do not see impacts on performance with multicollinear features (Gupta, 2021). At most, it can impact feature importance in the models so this will be something to keep an eye on later in the modeling.

**Methods, Testing, and Analysis**

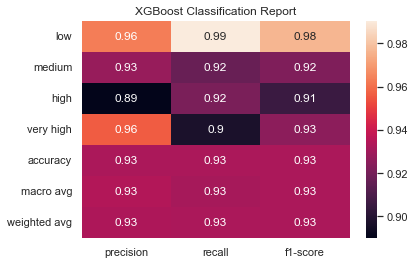
This project is trying to identify four different classes of price ranges. This makes this a multiclass classification problem. There are specific models that are designed to accept this type of problem. The two I focused on testing were Random Forest and XGBoost modeling.

Random forest is an ensemble model that uses individual decision trees to come up with class predictions (Yiu,2021). These decision trees map possible outcomes bases on a series of decisions that serve to identify what class something belongs to. In a random forest, multiple trees form decisions about the class an object belongs to and then the class with the most votes is chosen. There are a number of parameters that can be adjusted on these models to increase performance. I used a grid search to test different values for these parameters. The first baseline model without intervention on these parameters was 88% accurate in determine the right price grouping. After this, I employed the grid search to find the best parameters of the model. 6,250 model candidates were tested with three-fold cross validation. Three-fold cross validation is resampling procedure use to evaluate a mode with limited data (Brownlee, 2020). This validation process shuffles a dataset, then splits it into three groups, takes each group, and treats it as a holdout while using the other two groupings to train the model. So, each of the 6,250 models were fit three times, totaling 18,750 fits. After running these, the model with the largest improvement over the base model was chosen. The final model was 90% accurate, a 2% improvement over the base. Ram and battery power were the most impactful features on the model.

Chart

Description automatically generated with medium confidence

The second model that was tested was an XGBoost model. XGBoost stands for extreme gradient boosting. It is also an ensemble model, but each iteration attempts to minimize the errors of the last run (Morde, 2019). IBM describes it as “Boosting is an ensemble learning method that combines a set of weak learners into a strong learner to minimize training errors. In boosting, a random sample of data is selected, fitted with a model, and then trained sequentially—that is, each model tries to compensate for the weaknesses of its predecessor” (IBM Cloud Education). Once again, a baseline model was run that resulted in an accuracy score of 90%. After, a grid search was employed to find the optimal values for a set of model parameters. In this instance 3645 candidate models were run totaling 10,935 fits. Using the optimal parameters, the accuracy jumped to 93%. The presence of a touch screen was the most important feature in this model.



**Conclusion**

The best model was the XGBoost model that used the parameters with the best performance. This model was 93% accurate. By using three-fold validation, I was able to ensure that the model would perform well on a global population of data rather than just well on the sample data. This model can be implemented for use for any new phones that the company creates. We will need to work with IT to create a tool that allows users easy access to the model and come up with a plan to monitor the model for degradation. This information can then be used as a guide to help determine cell phone pricing. I foresee no ethical implications with this project so as long as company stakeholders find 93% accuracy acceptable, the implementation process can begin.

References

Brownlee, J. (2020, August 2). *A gentle introduction to k-fold cross-validation*. Machine Learning Mastery. Retrieved July 24, 2022, from https://machinelearningmastery.com/k-fold-cross-validation/

By: IBM Cloud Education. (n.d.). *What is boosting?* IBM. Retrieved July 24, 2022, from https://www.ibm.com/cloud/learn/boosting#:~:text=Boosting%20is%20an%20ensemble%20learning,the%20weaknesses%20of%20its%20predecessor

Gupta, V. (2021, February 2). *Correlation in xgboost*. Medium. Retrieved July 24, 2022, from https://vishesh-gupta.medium.com/correlation-in-xgboost-8afa649bd066

Morde, V. (2019, April 8). *XGBoost algorithm: Long may she reign!* Medium. Retrieved July 24, 2022, from https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

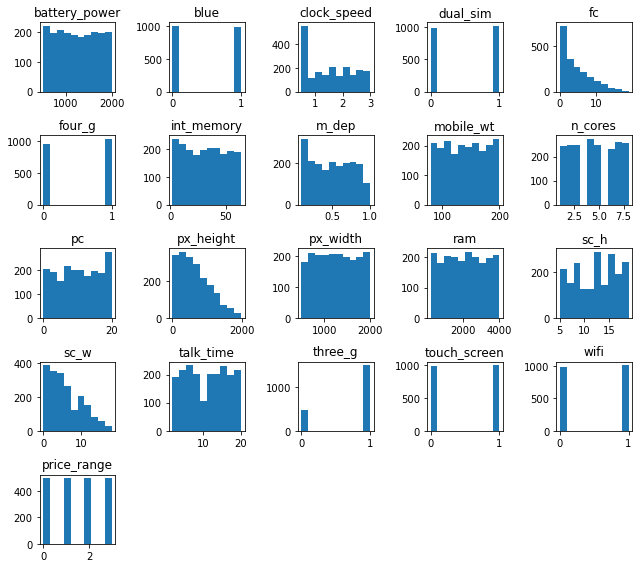
Pew Research Center. (2021, November 23). *Mobile fact sheet*. Pew Research Center: Internet, Science & Tech. Retrieved July 24, 2022, from https://www.pewresearch.org/internet/fact-sheet/mobile/

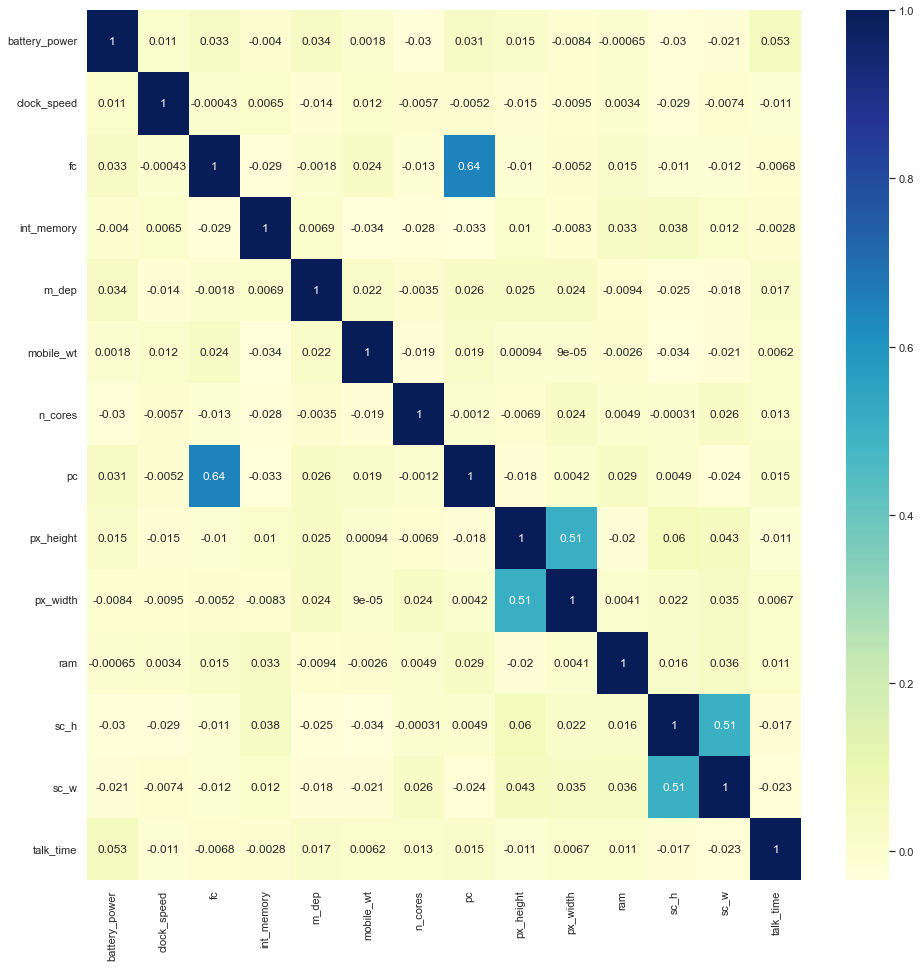
Raj, S. (2020, June 18). *Effects of multi-collinearity in logistic regression, SVM, RF*. Medium. Retrieved July 24, 2022, from https://medium.com/@raj5287/effects-of-multi-collinearity-in-logistic-regression-svm-rf-af6766d91f1b

Sharma, A. (2018, January 28). *Mobile price classification*. Kaggle. Retrieved July 24, 2022, from https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification

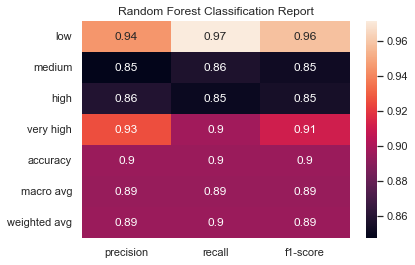
Yiu, T. (2021, September 29). *Understanding random forest*. Medium. Retrieved July 24, 2022, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Appendix

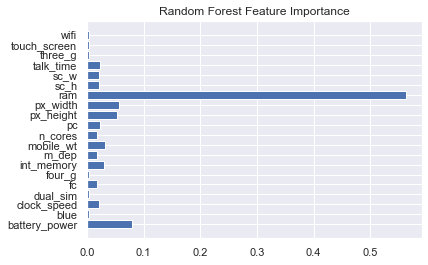
1.  Feature Distribution
2. Feature Correlation



1. Random Forest Results



1. Random Forest Feature Importance



1. XGBoost Results

Chart, funnel chart

Description automatically generated

1. XGBoost Feature Importance

