Abstract

We analyzed a data set of 200 DFCameras in an attempt to understand how certain features of the cameras impact their price and to also predict the price of new cameras. Our results indicate that depending on the needs of the physicist, a "good" amateur DFCamera (i.e. one that has a grade of amateur but is high in other features) may suffice because even a professional grade DFCamera with similar features will most likely cost more. As for manufacturers, we found that it is generally not profitable to further increase the maximum resolution of high-end cameras since professionals seem to care more about features such as portability and the number of resolution levels. Toward the predictions goal end, we created a regression tree using only 150 cameras, optimized the parameters of this tree based on its performance on a separate validation set of 50 cameras, and then use the optimized model to predict the prices of 200 new test cameras.

1 Introduction

Physicists on Tiaran use an equipment called a DFCamera to monitor the movements and changes of the dark energy fields that surround the planet. These cameras have a variety of features and characteristics, each with a range of values, that ultimately determine the price of a camera. In the most general sense, the price of a camera is determined by its grade (professional v. amateur), recharge time between shots, amount of energy needed per shot, angular coverage (how much of the sky can be seen), weight, maximum resolution, number of resolution levels, noise sensitivity, ability to connect to a network server, brand, ease of use, and release date. With these features in mind, we attempt to build a predictive model that, given the features of a new camera, can predict the camera's price (to some degree of accuracy). Such a model is useful not only from a prediction standpoint, but also to understand how each feature contributes to the selling price. Such information can then be used to answer interesting questions such as whether a good amateur DFCamera is worth buying over a professional camera with similar features, or whether it is worthwhile for manufacturers to further improve the maximum resolution on a DFCamera that is already high end.

To that end, we consider 200 randomly chosen cameras, whose features and price we know, to build a predictive model. We will then attempt to predict the prices of 200 additional randomly chosen cameras whose features (but not price) we know.

2 Exploratory Analysis

The data set we are given is called **DFC.train**, which contains 200 rows (corresponding to the 200 different cameras) and the column **PRICE** corresponding to the feature of the camera that we wish to predict, along with 12 columns **(GRADE, TIME, ENERGY, etc.)** representing the remaining features of the camera that will be inputs into the model. In this section, then, we explore the data via visualization and basic descriptive statistics to obtain a better grasp of what the data really looks like.

2.1 Response Domain

We generated the sample distribution of the response, Price (in 1000 Tiaran Dollars), in Figure 1 and it appears approximately normal. However, upon conducting the **shapiro.test**, we obtained a p-value of **3.361e-05**, which means that the prices of our sample cameras are not normally distributed.

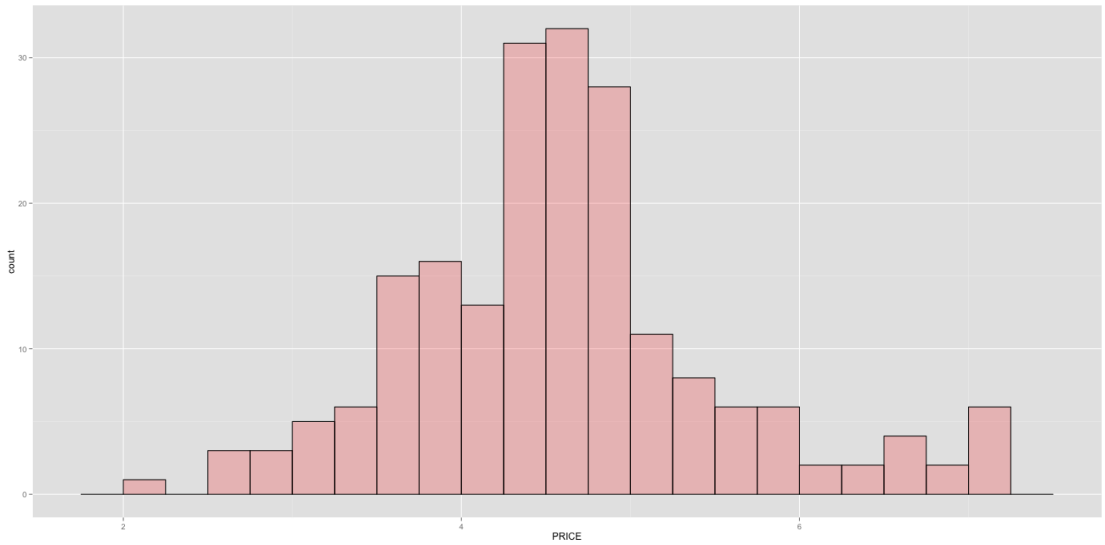

Figure 1: Histogram of Price (1000 TD)

2.2 Predictor Domain

For the predictor variables, exploratory analysis is a bit more tricky due to the fact that variables are of different types. For example, grade, brand, and ability to connect to network are factor variables (only take on a few discrete number of values) so their histograms do not show much interesting information. However, we found that noise sensitivity, angular coverage, and energy per shot seem bimodal, as in Figure 2, while the number of resolution levels is approximately normally distributed.

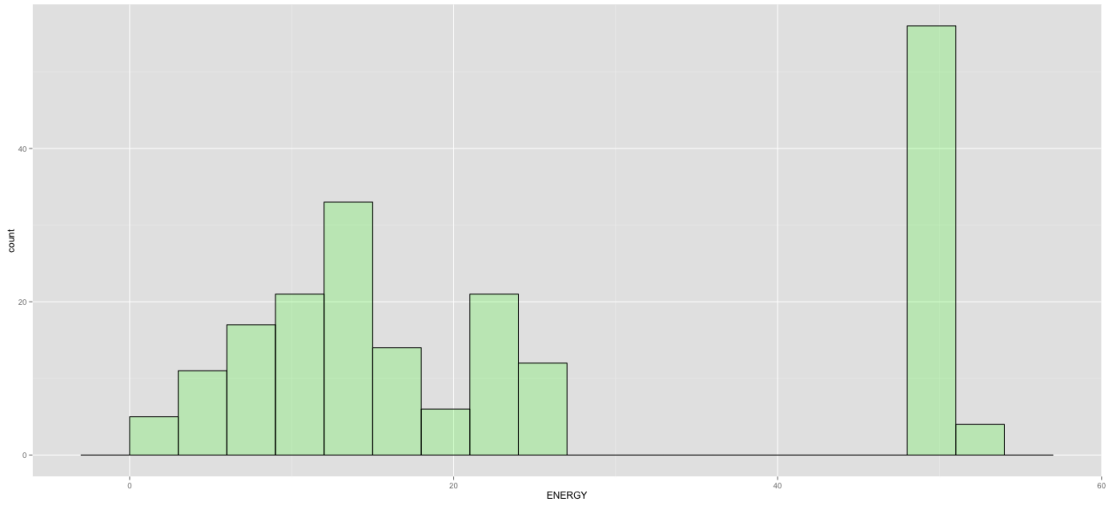

Figure 2: Histogram of Energy Per Shot (kWh)

However, we found that recharge time, weight, maximum resolution, ease of use, and manufacture date follow an approximately exponential distribution (having a very thin tail on either the right or left) as evidenced by, for example, Figure 3.

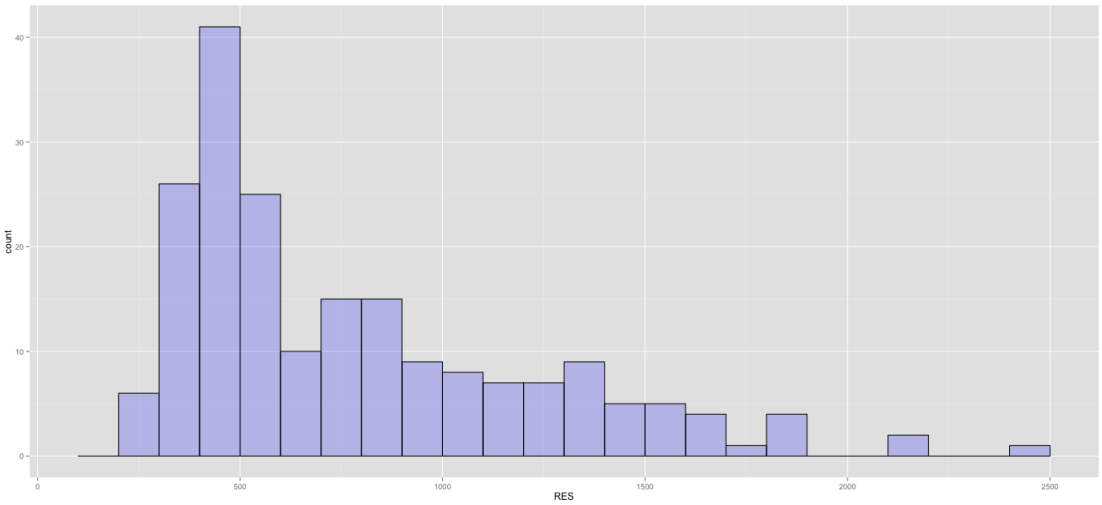

Figure 3: Histogram of Maximum Resolution (dpi)

Additionally, we discovered that some of these approximately exponentially distributed variables placed professional grade cameras at the tails and amateur grade cameras in the regions of higher density, which means that professional grade cameras are more rare but generally possess better features. Figure 4, for example, illustrates how professional grade cameras (blue dots) almost exclusively have short recharge times, while amateur grade cameras (black dots) possess recharge times across the entire spectrum.

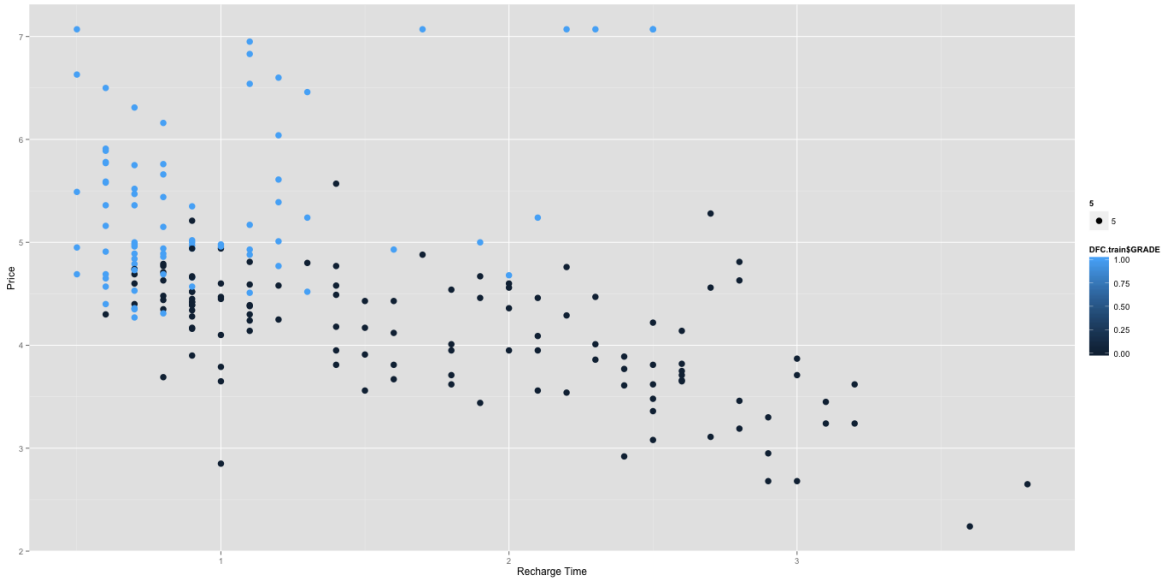

Figure 4: Recharge Time (sec) v. Price (1000 TD)

3 Methodology

3.1 Data Preparation

Before attempting to fit models on the data, some precautions had to first be taken. Firstly, we converted **GRADE, BRAND**, and **NET** into factors since these are qualitative variables.

Secondly, we standardized all of the numeric variables so that the size of the coefficients would be on the same scale. We standardized each variable X_i by the following method:

$$X_i^* = \frac{X_i - \bar{X}_i}{SD(X_i)} \tag{1}$$

where $\bar{X}_i$ denotes the average value of X_i and $SD(X_i)$ denotes the standard deviation (i.e. spread) of X_i. This metric hence converts the value of each variable into how far it is from its average value, in terms of how spread apart its values are. This means that variables that might be inherently larger than others (e.g. maximum resolution can be as big as 200 dpi while angular coverage is at most 8 degrees) can be comparable without worrying about units.

Lastly, we partitioned the data into our own training and validation set. This was achieved by randomly selecting 150 cameras from our sample to be the training set, and the remaining 50 to be the validation set. The purpose of this was to be able to construct the models using only the training cameras, but tuning the parameters of the models based on which models have the highest accuracy in predicting the price of the validation cameras.

3.2 Model Interpretation

From a model interpretation standpoint, an ordinary least squares approach would give us the advantage that the coefficients of each variable would provide us a quick and simple way to interpret the effect of each of the camera's features on its price. To this end, we carried out an all subsets regression in order to find the models that minimize AIC, BIC, and PRESS; three criteria commonly used to measure the goodness of fit of a model with multiple variables. Out of these three optimal models, we then chose the model that minimized prediction error on the validation set (smaller is better). We believe this methodology will allow us to find a model that removes variables that may already be correlated with other more important predictor variables, thus providing better answers to the research questions posed about the effect of specific features on camera price.

3.3 Prediction

While many of the camera's features were correlated with each other to some extent, this would not have an adverse affect on prediction as the range of our test set features mostly overlapped with the range of our training set features. However, since the data did not uphold normality assumptions, we believed an ordinary least squares approach would be suboptimal from a prediction standpoint. Indeed, given that features of higher-end cameras tend to cluster around high values (and vice-versa for lower-end cameras), a nonparametric model appeared to be the better choice for prediction. Ultimately, due to the limitations of our statistical expertise, we proceeded with the non-parameteric method most familiar to us: CART.

4 Results

4.1 Regression Performance

The optimal models under the AIC, BIC, and PRESS criteria (i.e. the models the minimized AIC, BIC, and PRESS, respectively) were all different, but the prediction accuracy on the validation set was greatest for the optimal model under the BIC criterion, which had only six variables in its model. For the CART, we used the same input variables as those from the best subset under the BIC criterion and found that the optimum validation error was achieved for a **minsplit** equal to 5.

Model	Variables	Validation Error
Minimum AIC	Grade Time Energy Wt Res ResL Sen Brand Date	0.5292592
Minimum PRESS	Grade Energy Wt Res ResL Sen Brand Date	0.5234532
Minimum BIC	Grade Wt Res ResL Sen Net	0.5143989
CART	Grade Wt Res ResL Sen Net	0.4019012

Figure 5: Model Performance

In particular, the ordinary least squares model using the variables from the optimal subset under the BIC criterion can be found in Figure 6.

Variables	Intercept	Grade	Wt	Res	ResL	Sen	Net
Coefficients	-0.36952	0.84122	-0.14819	- 0.27288	0.35937	- 0.33297	0.80393

Figure 6: Optimal BIC Model

This figure tells us that the biggest contributors to the price of a camera are whether it is professional grade, and whether it has the ability to connect to a network server. Additionally, cameras with a higher number of resolution levels, a lower sensitivity to noise, and a smaller weight all lead to higher prices, in general. Interestingly, the results suggest that cameras with a higher maximum resolution are, on average, cheaper.

Nevertheless, while the first three models in Figure 5 are useful for the purpose of understanding the effect that each predictor variable has on the price of a camera, from a prediction standpoint, the optimal regression tree achieves the minimum validation error. This means that the tree is the most accurate in predicting the prices of cameras that weren't even used to make the model itself. Additionally, the tree achieves this accuracy by using only six variables, which means it truly has a stronger predictive power.

4.2 Should Physicists Buy Amateur Grade Cameras?

The results conclusively showed that the feature of the camera that has the most influence on the price is its grade. Not only did **GRADE** have the largest coefficient in the standardized model chosen by the best subsets regression, but it was also the first variable to be included in forward stepwise regression. As such, it appears that when a camera is labeled professional grade, this increases the price significantly. Thus, we recommend that as long as a camera meets ones personal requirements for resolution, recharge time, noise sensitivity, etc. a physicist would be well off buying an amateur grade camera, as a professional grade camera with similar specifications would likely tack a premium onto the price.

4.3 Should Manufacturers Increase High-End Camera Resolution?

As for whether it is profitable for manufacturers to increase the maximum resolution of a high-end camera, the all subsets regression approach, while choosing the camera resolution among its best subset, generated a negative coefficient for resolution, indicating that an increase in resolution actually decreases the price. This implied negative association is not supported by the raw data, so the issue with this model is the presence of multicollinearity (i.e. the resolution is highly correlated with other variables in the model, masking its true effect).

Indeed, while resolution has a positive relationship with price, when we account for the grade of a camera, it turns out that the resolution of a professional grade camera is slightly negatively related to its price (a correlation of -0.297), as seen by the negative trend of the blue dots in Figure 7.

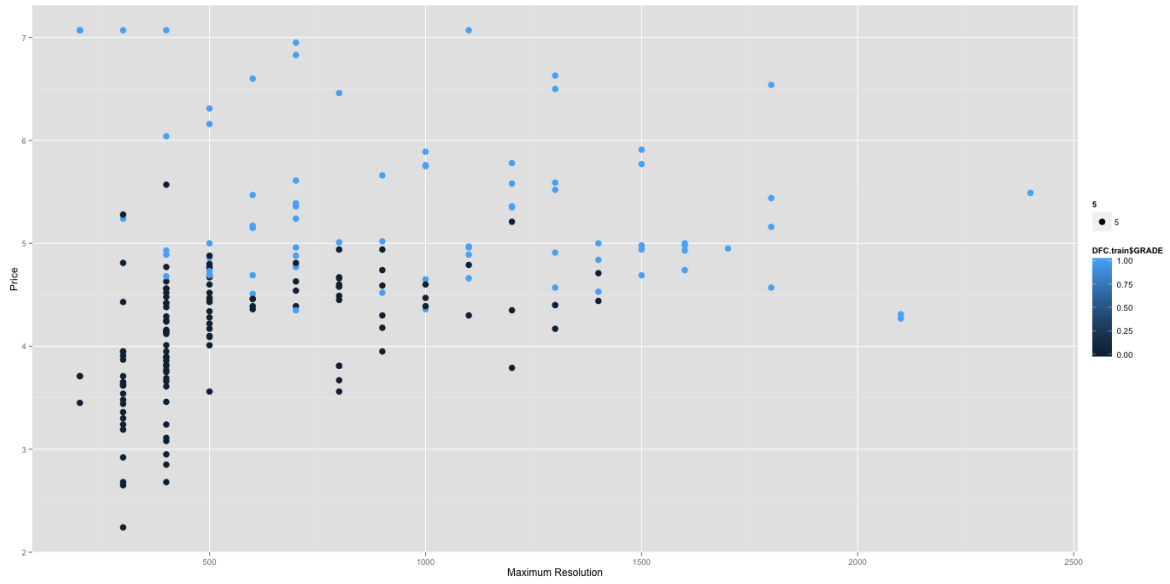

Figure 7: Maximum Resolution (dpi) v. Price (1000 TD)

At first glance, such a result is startling because we would expect a camera with a higher maximum resolution to be more valuable in general. It turns out, however, that this strange result is due to the fact that cameras with higher maximum resolutions tend to sacrifice other features, such as angular coverage.

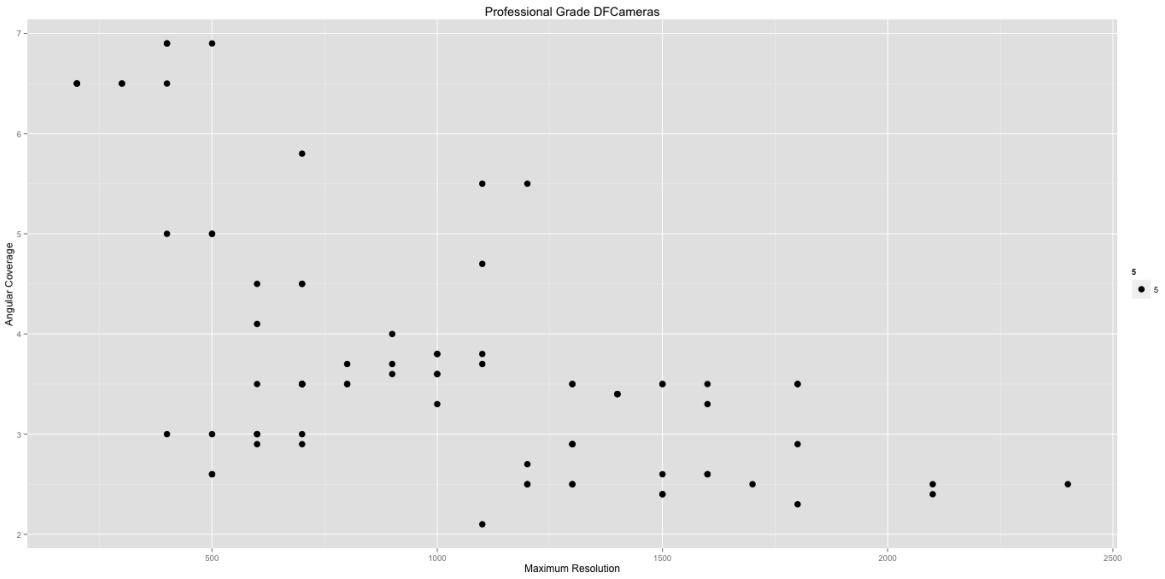

Figure 8: Maximum Resolution (dpi) v. Angular Coverage (degrees)

Additionally, it is likely that while amateurs care more about features like the maximum resolution, professionals buy the camera for other features such as angular coverage, number of resolution levels, etc. This is evidenced to some extent by Figure 9, containing the correlations between some of the features and price, for professional grade cameras.

Variable	WT	NET	ANGLE	RES	RESL
Correlation	-0.2031216	0.2877695	0.12767228	-0.296799952	0.53393861

Figure 9: Correlation Between Variables and Price (Professional Grade)

Indeed, if we assume that supply and demand are equal at current prices, this figure seems to suggest that professionals' preferences for portability (i.e. small weight), several resolution levels, angular coverage, and ability to connect to a network server outweigh their preferences for higher maximum resolution. As such, it appears that for high-end cameras, it would not be profitable to further increase the maximum resolution.

5 Discussion

One of the core questions in this research venture was to determine whether it is cost effective for manufacturers to increase the maximum resolution of the higher-end cameras. In order to answer this question, we needed to estimate the marginal increase in selling price of a high-end camera when the resolution was increased, given that other features of the camera are held constant. In general, cameras with higher resolution also tend to be of lower noise sensitivity while also having lower angular coverage, so isolating the effect of the resolution on the price alone is a difficult problem. Given that we barely scraped the surface of modeling in multiple dimensions, it is evident that we lacked adequate statistical expertise to attack this problem.

Additionally, in tackling the prediction problem, it would have been helpful if we had learned more nonparametric methods. Since this data was in higher dimensions, and many of the normality assumptions were violated, we were essentially limited to using regression trees. In the future, when enough statistical expertise is acquired, it may be more powerful to implement methods such as k-nearest neighbors and support vector machines, which may be useful for this type of problem in which clustering of features is highly likely.

Lastly, the finding that maximum resolution is negatively associated with price for high-end cameras was not fully explained. More work could have been done toward understanding exactly how this variable affects price, and it would likely involve exploring what effect a higher maximum resolution has on other desirable features of a camera. While we discovered a negative association between maximum resolution and angular coverage, this association is surely not exhaustive. As such, it may be worth exploring other specifications of the cameras that were not included in our data set, which could be achieved by surveying customers who buy high-end cameras to understand what features are more important to them.