Abstract

Welcome to my Biomedical Data Science Capstone Project where I analyze the multivariate chemical composition of wines in a standardized data set so that I can understand the classification structure of wine and test if a computational model can be used to predict the cultivar of grape used in the creation of the drink. After I conducted data quality checks for my set, I used correlational analysis and Principal Component Analysis(PCA) to visualize the multidimensional relation between the physicochemical variables that could influence grape cultivar prediction. I then trained and evaluated two complementary classifiers: a multinomial logistic regression and a random forest to evaluate if a relatively simple data-driven model could be used to predict future wine classification.

Introduction

Classification of biological and chemical samples with measurable attributes is a hallmark of biomedical data science and as such I decided to pick the wine data set because it seemed to have a large amount of variety in the data that it had attributed to each wine sample. Features were acquired from standard chemical assays(alcohol, acids, phenolics, color intensity, etc.) and the associated cultivar labels were also known according to the documentation. The end goal of this project was two-fold. I first wanted to understand the multivariate structure of the data set in the chemical space to try and gain insight into whether or not the cultivars are separable when all features act and are used together (in weighted proportion of course) for prediction. Upon success, I then wanted to see if these insights could be used to test predictability out-of-sample using complementary models that are good for balancing interpretability and flexibility.

Data and Methods

Data Quality Checks: The first thing that had to be done was making sure that the data was of good quality. This included checking if there were missing values, incorrect data types, and a reasonable spread across features (an actually useful/interesting data set). These attributes are shown in Table 1 & 2. This was quite easily done in the code using built in package functions. Making Figure 1 further proved to me some sort of structure existed to tap into for predictive purposes.

Exploratory Structure: The first thing that I did after verification that the project was worth pursuing was generate primitive data characteristics (Table 3) and create a correlational heatmap (Figure 2). From this I wanted to see if any of the physicochemical markers were highly tied with each other. One thing to note here was that the variables were not independent of each other and there were some areas of “heat” in the heat map that suggested high correlation between different variables. This seemed to suggest a sort of redundancy or multicollinearity in the data set and so I had to keep this in mind if I wanted to predict the classification of wine. Given this, I then decided to conduct a PCA analysis. After standardization PCA projects each wine into two orthogonal directions that capture the largest joint variance across all the features. Figure 3 shows visible clustering indicating that there does indeed exist an underlying class structure at the multivariate level. Figure 4 shows some of the key informative features plotted against each other because seeing the clustering on the PCA I decided to look back at the heatmap for any highly correlated data attributes and realized that plotting the areas with moire “heat” as mentioned before was probably most easiest when it came to visualizing what variables may be acting multicollinearly to influence cultivar category. This satisfies a key precondition for supervised learning and so was an exciting development because it allowed me and gave me the basis to move on, actually creating a supervised learning model.

Supervised Learning and Evaluation: The first step that was conducted was receiving 20% of the data set as a holdout test set with stratification so that class proportions could be preserved. The baseline model that I made was a multinomial logistic regression with standardized inputs and the nonlinear benchmark I then made was a random forest algorithm that is said to better be able to capture the nonlinear relations in the data. Once this was completed I then utilized 5-fold cross validation (Table 4) and confusion matrices (Figure 5) to easily visualize the validity of the models I had made. Once the random forest training and testing were done I used the calculations done there to output Figure 6 which shows the features that were the most influential in determining cultivar category and is ideally the insights that I would want to show others in the community if they desired to determine their wine's grape cultivar based on tangible lab measurements rather than just word of mouth. The reason I decided to make and validate both models was because I wanted to see if the linear assumption in the logistic model sufficed, which it nearly did here, as well as quantifying potential gains from modeling interactions and from using one model over another.

Results

As it can be seen in Table 4 the CV accuracy was 0.993 ± 0.014 and the hold out accuracy was 0.97. The test classification report also shows high precision and recall across classes (notably, class_1 recall = 1.00; class_2 recall = 0.90). For the random forest it can be seen in the notebook that the hold out accuracy was 1.000 showing that all 36 test samples were correctly classified in this split.

In terms of model interpretability it makes sense to explain the logistic regression coefficients and the random forest feature importances. Logistic regression coefficients indicate how increasing a feature (by 1 SD) shifts the log-odds toward a class. Positively weighted features for a class help delineate cultivar categories. In practice though, features associated with phenolic content (e.g., flavonoids, total phenols, OD280/OD315) and color intensity/proline strongly influence the boundaries—consistent with the PCA loadings intuition and chemical expectations. Meanwhile, the Random Forest bar chart (Figure 6) highlights similar chemistry (color_intensity, proline, flavanoids/phenolics) as top contributors, providing a black-box corroboration of the Logistic Regression story while allowing for nonlinear thresholding and interactions.

Discussion

There were a few things that I learned through this data analysis. First, I learned and implemented multivariate separability with PCA which showed the distinct clustering of data points that led me to train my model. It implied that combined chemical signatures and not just a standalone feature drove the separability of the wine cultivars. I also learned about predictability. Both the Logistic Regression and the Random Forest performed strongly when out-of-sample, i.e. the test data, was provided and so this helped me see how visual structures and representations could be used for real predictive consequences. The final thing I learned about was convergent interpretability and multimodal validation. Logistic Regression coefficients and Random Forest importances emphasized a consistent chemical subset in the data (phenolics/flavanoids, color intensity, proline) showing how model evidence aligned with domain intuition.

Limitations and the Future

Given that this is a data analysis it also makes sense to discuss some of the shortcomings of the data set and any future steps I might take if I were to create a more thorough analysis. First the wine data set is a small clean and curated set that isn't very much like a real biomedical data set which is typically larger and noisier with outliers and confounds. Secondly, the features here are well behaved continuous laboratory measurements so there are no mixed data types, batch effects, time dependencies, or measurement drift effects which are all things that are usually quite common in actual biomedical data. Finally, the data set is a single-source set with no external cohort (hence why I had to train-test split my one set of data rather than training with one set and testing on a different set that has the same measurements but is from a different lab or what not). In this way it is impossible to assess generalization beyond this specific collection.

Beyond the data itself, there are also some limitations in the analysis I did. For one thing only the logistic regression model was validated with cross-validation, whereas the random forest was only validated on a single hold-out split, which risks optimistic or pessimistic bias depending on the random partition. Evaluation also used accuracy as the primary metric which is good for just a single data set but across multiple data sets is not necessarily the best metric to use. It does not examine the probability of some classes being more error prone than others and is sort of archaic in that it only examines right and wrong without really trying to further consider the reason for why it would be right and wrong. This could be a place to use a gradient boosting algorithm where eros can be fed back to the model and investigated further but given how much data I had, it would be difficult to curate more data to add to the training/testing set.

This segues naturally to what future analysis might be done and as stated I could incorporate more cross-validated benchmarking through SVM, gradient boosting, etc. I can also work harder to incorporate interpretability tools such as particle dependence or SHAP to better articulate any feature interactions captured by the forest. Robustness checks across random seeds, class-balanced splits, permutation tests, and sensitivity analyses would reduce the risk of overfitting into a particular partition.

References

1. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. *JMLR*.
2. Loader dataset: sklearn.datasets.load_wine() (UCI origin; chemical analysis of wines).
3. Jolliffe & Cadima (2016). Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A*.
4. DS4Bio Book: *Data Science and AI for Bio/medical applications using Python* (smart-stats.github.io/ds4bio_book).

Figure 1: Preliminary Exploratory Graphs

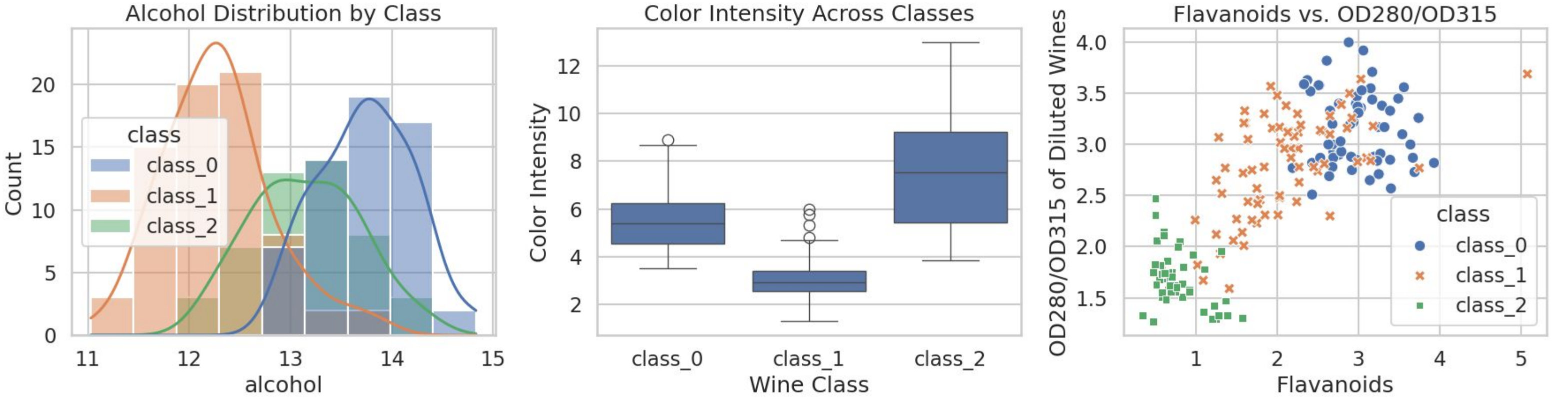


Figure 2: Correlational Heat Map

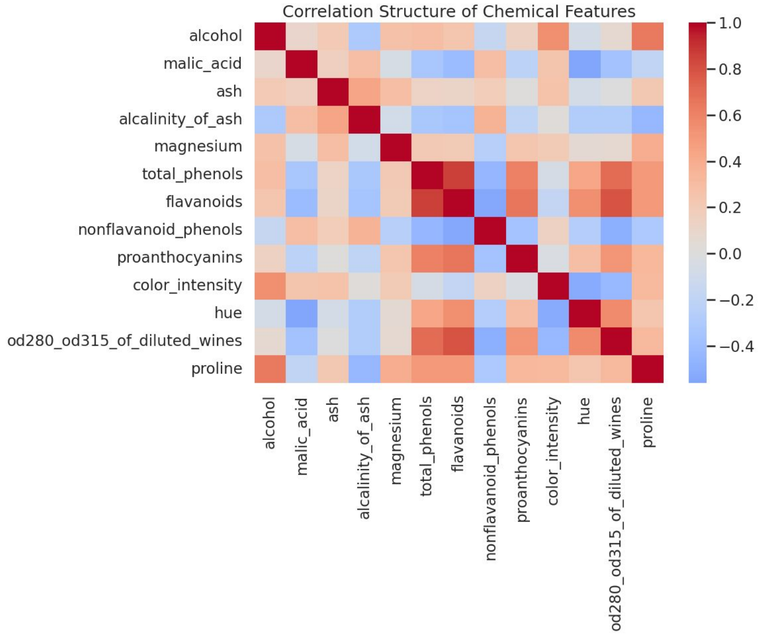


Figure 3: PCA Plot

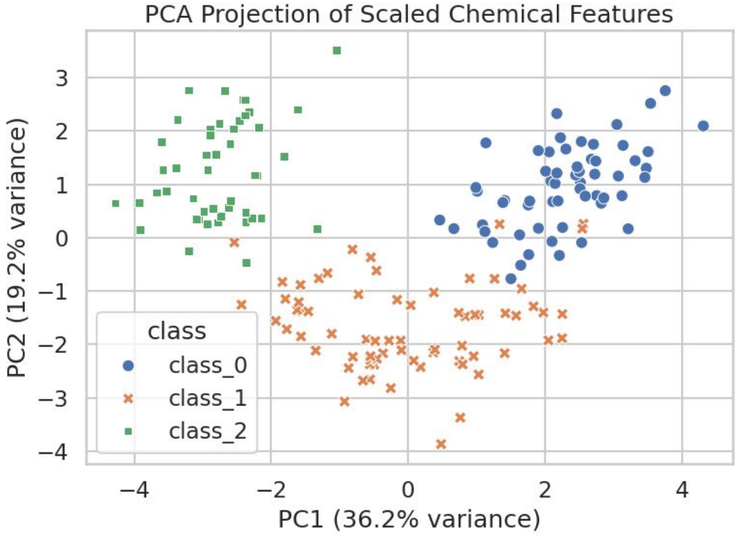


Figure 4: Correlation For Features of Interest

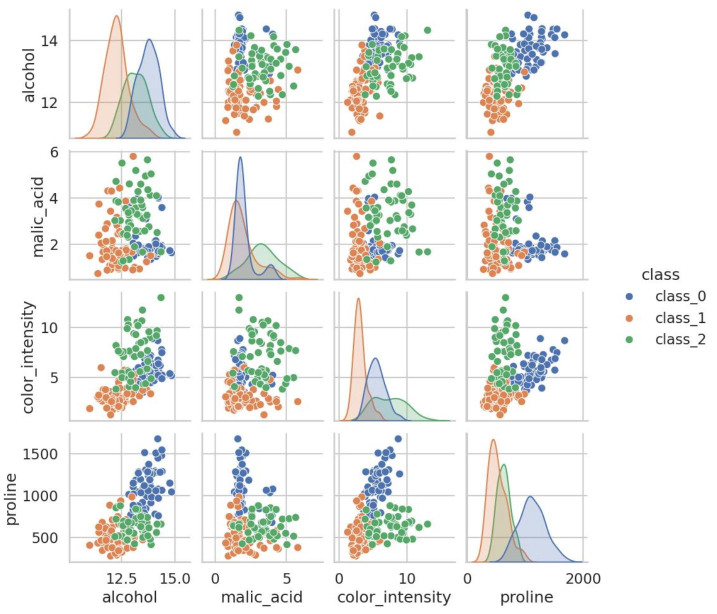


Figure 5: Model Confusion Matrices

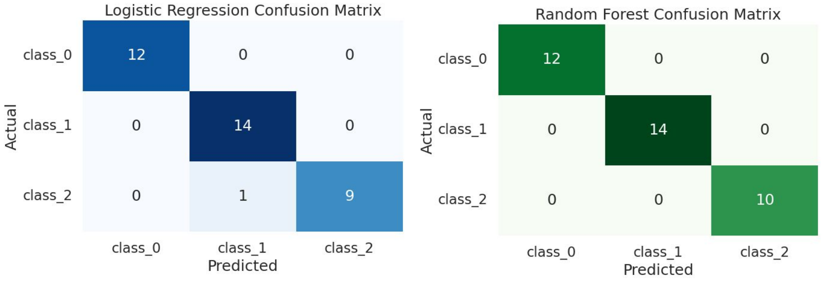
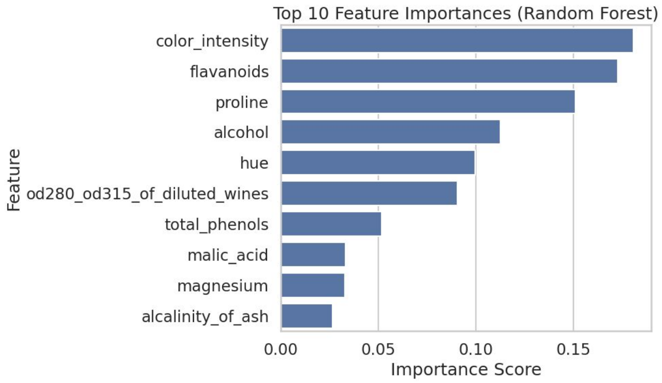
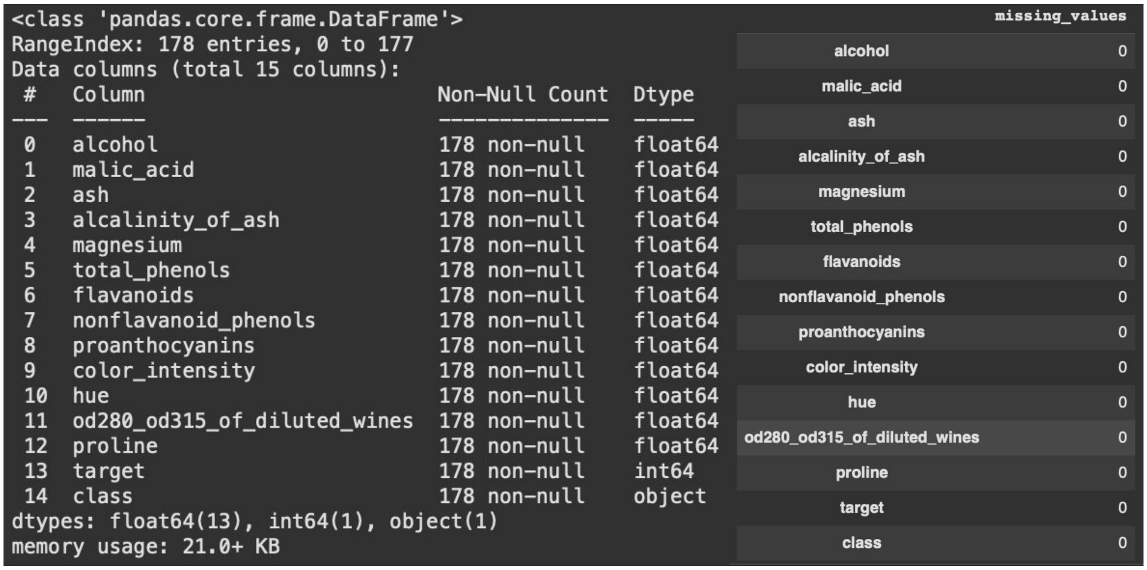


Figure 6: Feature Importance

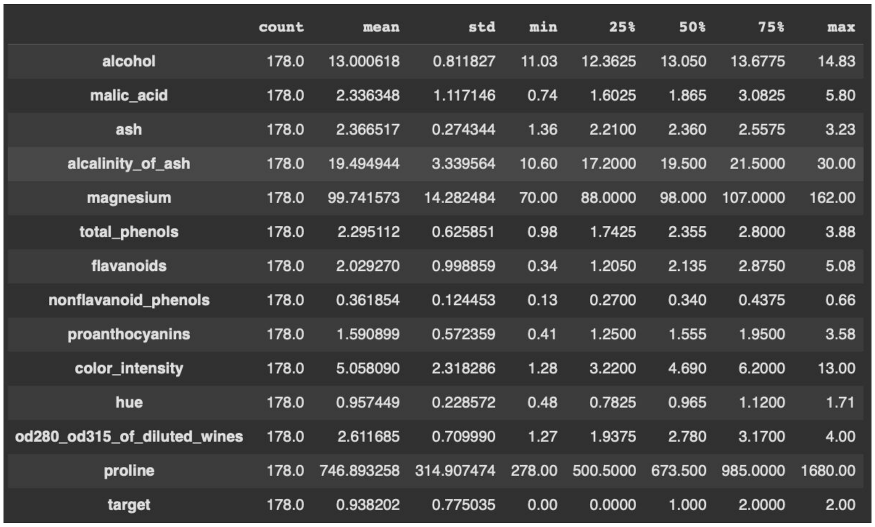




```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 178 entries, 0 to 177
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   alcohol                       178 non-null    float64
 1   malic_acid                    178 non-null    float64
 2   ash                           178 non-null    float64
 3   alcalinity_of_ash             178 non-null    float64
 4   magnesium                     178 non-null    float64
 5   total_phenols                 178 non-null    float64
 6   flavanoids                    178 non-null    float64
 7   nonflavanoid_phenols          178 non-null    float64
 8   proanthocyanins               178 non-null    float64
 9   color_intensity               178 non-null    float64
 10  hue                           178 non-null    float64
 11  od280_od315_of_diluted_wines  178 non-null    float64
 12  proline                       178 non-null    float64
 13  target                        178 non-null    int64
 14  class                         178 non-null    object
dtypes: float64(13), int64(1), object(1)
memory usage: 21.0+ KB
```

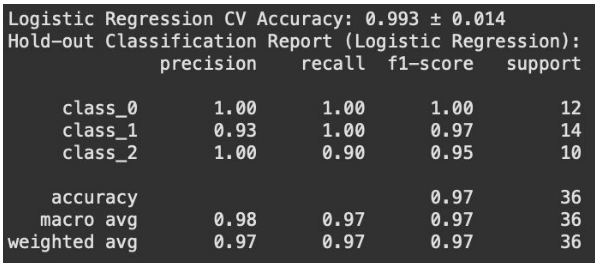
	missing_values
alcohol	0
malic_acid	0
ash	0
alcalinity_of_ash	0
magnesium	0
total_phenols	0
flavanoids	0
nonflavanoid_phenols	0
proanthocyanins	0
color_intensity	0
hue	0
od280_od315_of_diluted_wines	0
proline	0
target	0
class	0

Tables 1 & 2: Shows data quality check outputs for count, data type and missing values



	count	mean	std	min	25%	50%	75%	max
alcohol	178.0	13.000618	0.811827	11.03	12.3625	13.050	13.6775	14.83
malic_acid	178.0	2.336348	1.117146	0.74	1.6025	1.865	3.0825	5.80
ash	178.0	2.366517	0.274344	1.36	2.2100	2.360	2.5575	3.23
alcalinity_of_ash	178.0	19.494944	3.339564	10.60	17.2000	19.500	21.5000	30.00
magnesium	178.0	99.741573	14.282484	70.00	88.0000	98.000	107.0000	162.00
total_phenols	178.0	2.295112	0.625851	0.98	1.7425	2.355	2.8000	3.88
flavanoids	178.0	2.029270	0.998859	0.34	1.2050	2.135	2.8750	5.08
nonflavanoid_phenols	178.0	0.361854	0.124453	0.13	0.2700	0.340	0.4375	0.66
proanthocyanins	178.0	1.590899	0.572359	0.41	1.2500	1.555	1.9500	3.58
color_intensity	178.0	5.058090	2.318286	1.28	3.2200	4.690	6.2000	13.00
hue	178.0	0.957449	0.228572	0.48	0.7825	0.965	1.1200	1.71
od280_od315_of_diluted_wines	178.0	2.611685	0.709990	1.27	1.9375	2.780	3.1700	4.00
proline	178.0	746.893258	314.907474	278.00	500.5000	673.500	985.0000	1680.00
target	178.0	0.938202	0.775035	0.00	0.0000	1.000	2.0000	2.00

Table 3: Baseline statistics of physicochemical wine markers

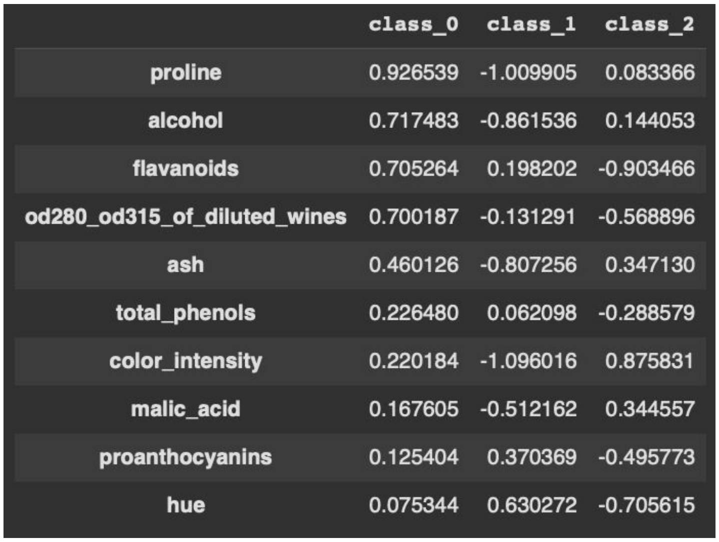


```
Logistic Regression CV Accuracy: 0.993 ± 0.014
Hold-out Classification Report (Logistic Regression):
              precision    recall  f1-score   support

     class_0       1.00      1.00      1.00        12
     class_1       0.93      1.00      0.97        14
     class_2       1.00      0.90      0.95        10

    accuracy                           0.97        36
   macro avg       0.98      0.97      0.97        36
weighted avg       0.97      0.97      0.97        36
```

Table 4: Logistics Regression Cross Validation Accuracy



	class_0	class_1	class_2
proline	0.926539	-1.009905	0.083366
alcohol	0.717483	-0.861536	0.144053
flavanoids	0.705264	0.198202	-0.903466
od280_od315_of_diluted_wines	0.700187	-0.131291	-0.568896
ash	0.460126	-0.807256	0.347130
total_phenols	0.226480	0.062098	-0.288579
color_intensity	0.220184	-1.096016	0.875831
malic_acid	0.167605	-0.512162	0.344557
proanthocyanins	0.125404	0.370369	-0.495773
hue	0.075344	0.630272	-0.705615

Table 5: Random Forest Coefficients