Final Project

Ziad Abouchadi, Susann Almasi, Aditya Bindal, Jinesh Ramani

December 11, 2015

1 Introduction

1.1 Overview

The market of retail diamonds is highly competitive. As dimaonds are a commodity, a variety of categorical and numerical variables influence their price. Companies in this space often use competitor prices as a benchmark when pricing diamonds.

As a new online retailer for diamonds, we will study the listed inventory of our competitors to improve our pricing and purchasing decisions. Specifically, we have two objectives:

1. Predict the price of a diamond based on various factors to develop pricing formulas for our diamonds. This results from this model will help us ensure that our prices are competitive.
2. Identify large clusters of diamonds in our competitors' inventory to determine our wholesale purchasing strategy.

To answer both of these questions, we will employ supervised and unsupervised learning techniques.

1.2 Data

Using Ruby, we built scrapers to collect data on diamonds listed on Brilliant Earth, an online jewelry retailer. By identifying Brilliant Earth's XHR, we were able to access structured data on the retailer's inventory, yielding the following for each diamond:

```
{"origin": "Botswana DTC", "symmetry": "Excellent", "suggestions":
"1699301A\n0.30 Carat Round Diamond\n\n", "shipping_day": 6, "report": "GIA",
"shape": "Round", "length_width_ratio": 1.0, "polish": "Very Good", "clarity":
"SI2", "id": 1798962, "title_s": "0.30 Carat Round Diamond", "cut": "Ideal",
"orderby_short": "2 PM PT monday", "title": "0.30 Carat Round Diamond",
"clarity_order": 1, "measurements": "4.31 x 4.28 x 2.66", "carat": 0.3,
"length": 4.295, "color": "E", "valid": true, "receiveby_short": "Tue, Dec 15",
"supplier": "Dharm", "_version_": 1519620841350889475, "product_class_exact":
"Loose Diamonds", "rap_percent": -47.0, "report_order": 4, "shipping_supplier":
"Dharm", "price": 550, "collection": "", "price_exact": "550.0", "culet":
"None", "active": true, "table": 59.0, "orderby": "Monday December 7, 2015 by
2:00 PM PT", "color_order": 6, "fluorescence": "None", "girdle": "Medium -
Slightly Thick", "cut_order": 5, "upc": "1699301A", "depth": 61.9, "is_memo":
false, "be_price": 302.1, "receiveby": "Tuesday, December 15"}
```

By combining these JSON data, we created a clean dataset of `r` diamonds. Our data contain `r` factors, excluding attributes that we are not using (e.g., SKU ID, UPC etc.)

Appendix B contains an sample of our input data.

2 Predicting Price

2.1 KNN

2.2 Trees

2.3 Random Forests

2.4 Boosting

3 Finding Clusters

3.1 K-Means Clustering

4 Conclusion & Recommendations