Couture Predictor 1 Report
RF Regressor on luxury clothing brand data

Myarion Johnson

March 2025

Abstract

This report explains the process of carrying out the Couture Predictor 1 ML experiment. In this initial experiment, a regression model is created to predict the price of Prada's clothing across five collection categories. It consists of a hypothesis of the model capability expectation, the method of data engineering, the preparation of the experiment, and the concluding results.

The combined training and testing data set consists of 288 rows and 4 columns: category, material, description, and details. The features' correlation to each other and the price is measured and displayed. This allows an overview of feature significance and classification ability. The insight on the predictor variables' usefulness gives an inference on how the model will perform.

The role of this experiment is to understand the pricing pattern in luxury clothing. After accurately transforming data into a feasible form for machine learning models, the value of the features from unknown or random chance residuals from price predictions can be differentiated. In further experimentation, advanced features such as images, may be used with image classification models to uncover more patterns not identified from text features. Future experiments may also include more data, including more categories and data across more brands, which can be used to improve the model's accuracy.

Introduction

When considering a product's price, you typically consider its components: what it is made up of and their value. Like food with ingredients and a method of being prepared; clothing is made of materials- manufactured and designed in a certain way. In contrast, a product's price may be heavily influenced by its brand. By definition, luxury is having great comfort and extravagant living provided by **expensive** and beautiful things. This gives luxury brands the ability to price products beyond their constructed value. The invisible value increase for whatever reason it may be post-production, will take extra intuition from a machine learning model to detect. Running a statistical experiment on data from many examples will allow you to dissect the available information involved in the manufacturing of products and understand the value they give to them. The residual here, less likely to be "random chance" may be categorized into things like trends (popularity- brand or product related), advertising, or other strategic but undocumented processes involved with the item's retail.

Hypothesis

The hypothesis constructed for this experiment is direct to open the door for this research: Prada's clothing prices can be predicted by its category, material, description, and details.

Data Collection

Using Python and a web scraper library, Selenium, the data of all site-available products across 5 different categories were scraped for a total of 288 clothing examples. This consisted of 60 outerwear pieces, 63 denim, 42 leather clothing, 110 knitwear, and 13 suits. The formatting of the material came as a single material (presumptively 100%) or the percent of each material used to create the product. Using a simple function to extract the information from values into variables, a list of unique materials was created to allow an embedding transformation for the column. An encoding was created, similar to one hot encoding, using the percent of each product's material as a float from the range 0.00 to 1.0 to represent said material's presence and weight in the clothing. The description- a paragraph of text, and details- list of text could be transformed using the spaCy natural language processing library. Specifically, the "en core web lg" model was used to convert the features to vector space, to be usable as input variables. The normalized vector was applied to created a new column that would represent their feature values. The more simple feature transformations applied were label encoding of clothing categories and price conversion to float. After appropriately cleaning and converting the data, we could document correlations to observe feature-effectiveness prior to any training.

Preparing The Model

The selected features- description, details, category, and material are split into training and test batches with a 4:1 ratio (80%/20%). The scikit library will be or toolkit for creating and evaluating the model. We create 15 random forest regressor (RFG) models to explore the parameters n-estimators [50, 100, 200, 300, 350] and max-depth [None, 1, 2]. Each RFG is evaluated with the mean absolute error and R-squared metrics.

Results

The best performing model of the 15 has a mean absolute error of 1243, and an R-squared score of 0.4986. This means the predictions were off by about $1.2k, and the variability could be about half-explained by the model and predictors. The price column's values fluctuated between the clothing, ranging from $1100 to $23900 with a mean of $3554 and a standard deviation of $2834. That residual could possibly be explained as a "definition of luxury" or "trend in fashion", being the offset of pricing beyond the manufactured value of the product as a regular piece of clothing. A preferred graphing library was used to visualize the correlations in the stage of understanding the data, in this case, plotly expressed. The scatterplot in figure 1 at the end of this report presented little to no patterns with price. This would, on the surface at least, indicate poor statistical significance of these two features on the target label. In contrast, when using boxplots as seen in figures 2 and 3, there was clear variation between each category, indicating the possibility to classify clothing types using these three features.

Conclusion

We hypothesized that Prada's clothing price can be predicted by its category, material, description and details using the present data on their web page. The data was gathered and transformed to be compatible with machine learning models, as well as comparable to each other for relationship analysis. Collecting results from this analysis and metrics on the regressor model, the features determined themselves capable of classification, but still leaving residual during target prediction. This experiment presented insight on buyer-side information predictability and can be succeeded with experimentation on more data across multiple luxury brands where a synthetic feature of brand popularity may construct itself, and the clothing image feature for style and pattern classifying.

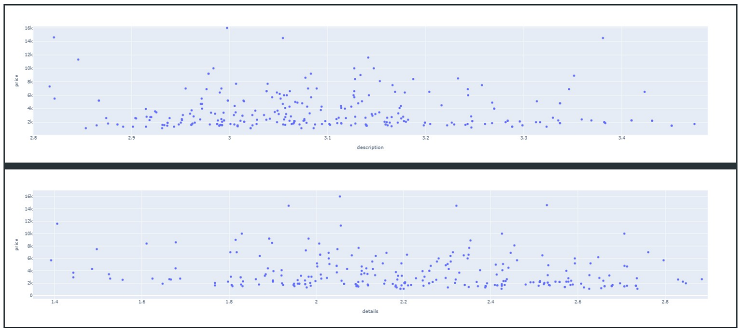

Figure 1: Scatterplot of description(top) and details correlation to price

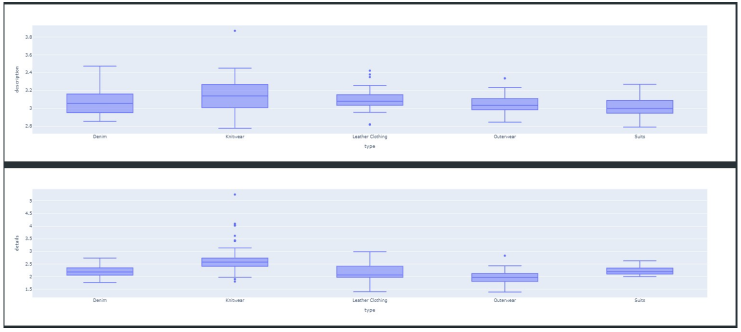

Figure 2: Boxplot of description(top) and details correlation to categories

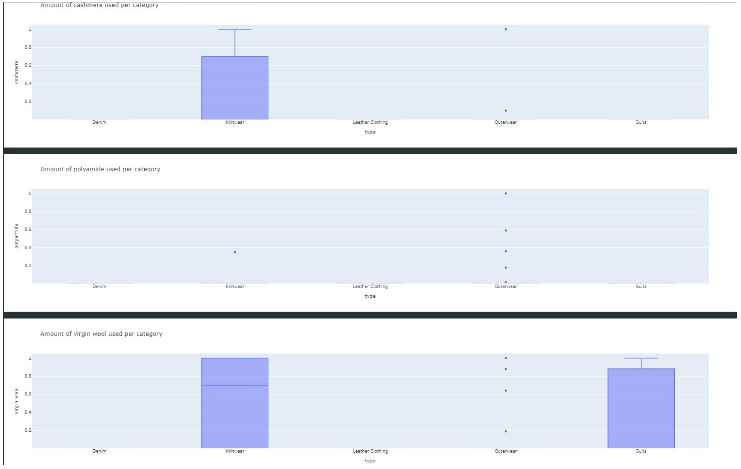

Figure 3: Boxplot of 3 example materials' correlation to categories