Diamond price estimate using numeric and visual data

Abstract

The main goal of this project is to correctly predict the price of a diamond using either numeric data or visual data.
The first model we tested was a simple Linear regression model which we fed numeric data.
The second model we used was that of a Neural Network using 132 neurons in the hidden layer.
Between the two, the first linear regression model gave the better results. The third model we tested was that of CNN using images of diamonds in various shapes.

Introduction

Evaluating a diamond is a difficult job usually left for the professionals. In this project we will test if by using raw data and machine learning we could correctly predict the value of a diamond. The value of a diamond is usually mostly affected by the four big C's (Carat, Cut, Clarity, Color) each has his own grading scale. By converting the grading into numerical data, we will be attempting to see how valuable those features really are. In addition there are other important features such as dimensions, symmetry and fluorescence which is the sparkle sometimes visible in certain diamonds.

Related work

There have been many papers written about deep learning and price predicting of various objects. The main subject which is also one of the most profitable niches is stock price predictions. Many companies are built on deep learning algorithms in the specific goal of correctly predicting the stock price.

Required background

To better understand the rest of this article we will be going over some terminology. Firstly, the most basic model for data processing, Linear regression. Linear regression is a linear approach

As we said this project presents many different models when each of them works in a different method of data processing.
The first and most basic model is linear regression. Linear regression is a linear

approach to modelling the relationship between a scalar and one or more variables (dependent/independent variables). Numeric features are calculated linearly to correctly guess the scalar. In the case of our project, it is to predict the price of a diamond.

The second model, Neural Network, is essentially a series of algorithms that try to recognize underlying relationships between variables(features) through a process that mimics how a human brain would work. It can decide which variables affect each other and create a better understanding for which variables are dependent or independent.

Lastly CNN – Convolutional Neural Network is a network that specializes in processing data that has a grid-like topology, such as an image in our case. Could also be useful with voice recognition and video.

Experiments and Results

Using basic linear regression gave surprisingly good results in compare to its simplicity. With a r^2 score of 0.83 it was a tough act to follow.

On the first attempt of creating the NN, we used the sigmoid function for our loss and normalized the label in order to get the values in the range of 0-1. This gave bad result with a loss of over 1,000,000. After multiple attempts, the combination of decreasing the learning rate, using the reduced mean loss function and not normalizing the data gave the best results with a loss of about 15,000.

In sharp contrast to the first two attempts in which only numeric data was used. With the CNN approach we used images of diamonds instead. The most difficult aspect was processing the data since not all the diamonds in the database had usable images. At first we downsized the images to 130x130 and ran a large batch size. The results were not impressive. With high loss, high RMSE and negative R^2 score. Once we used the full size image 300x300, increased the epochs, shrank the batch size and shrank the kernel size, the results improved. Reaching a loss of 0.002 and RMSE of 0.05. The only issue we couldn't address was the R^2 Score which stayed in the negatives (-3). This means that compared to the standard model, our CNN model was making worse predictions.

```
21/21 [==============================] - 40s 2s/step - loss: 0.0040 - val_loss: 0.0028
Epoch 18/20
21/21 [==============================] - 40s 2s/step - loss: 0.0040 - val_loss: 0.0025
Epoch 19/20
21/21 [==============================] - 40s 2s/step - loss: 0.0041 - val_loss: 0.0022
Epoch 20/20
21/21 [==============================] - 40s 2s/step - loss: 0.0038 - val_loss: 0.0022
Tesst RMSE:0.05755
Test R^2 Score: -3.47714
```

Project description

Preprocessing-

- Converted all textual grading values into numeric values(EX,VG,GD… -> 5,4,3,…)
- Removed unnecessary features(url, shape)
- Split single complex features into multiple features(HxWxD -> H , W, D)
- Sorted out all unusable images

CNN:

Matched filepaths with viable images and loaded them, separated the images into train/test/validate + normalized for rgb colors. Shuffled the data, created more features and then flattened the data so it could be processed.

Conclusions:

To conclude, each model we used had their own benefits. The linear regression gave the best R^2 Results. CNN gave the best loss results and the NN was average. In order to use CNN more affectively in this instance, it is necessary to have higher quality images and images from multiple angles. Diamonds have mostly straight forward pricing therefore if we are price predicting by numeric data, linear regression is the best option.