Project-Prompt

January 10, 2022

1 The Diamond Challenge: Project Prompt

2 Your new job

Congratulations! You've been hired as a data scientist at a diamond wholesaler. The retailer has collected data on diamond sales over the past 10 years. Your company would like to better understand the diamond market, identify trends, and build a model that can predict diamond prices for next year. You will put your data analysis skills to the test by uncovering insights about the diamond market.

This notebook describes the key deliverables you will be responsible for. It is recommended (but not required) that you use data science languages/tools such as Python or R for your work. Or a combination of many tools depending on the task at hand.

This Jupyter notebook introduces your new job assignment in Python. You can view the notebook on Github, or download and run it for yourself.

3 The dataset

You will use the dataset wholesale_diamonds.csv, which the company has compiled over the past 10 years. Lets take a look.

[1]:
```
# load the datafile and inspect the first 5 entries. Each row is a diamond
import pandas as pd
df = pd.read_csv("wholesale_diamonds.csv")

print("Number of rows = %i"%len(df))
df.head()
```

```
Number of rows = 407280
```

[1]:

	index	carat	cut	color	clarity	depth	table	cost (dollars)
0	0	0.23	Ideal	E	SI2	61.5	55.0	326
1	1	0.23	Good	E	VS1	56.9	65.0	327
2	2	0.29	Premium	I	VS2	62.4	58.0	334
3	3	0.31	Good	J	SI2	63.3	58.0	335
4	4	0.24	Very Good	J	VVS2	62.8	57.0	336

	length (mm)	width (mm)	height (mm)	year
0	3.95	3.98	2.43	2010
1	4.05	4.07	2.31	2010
2	4.20	4.23	2.63	2010
3	4.34	4.35	2.75	2010
4	3.94	3.96	2.48	2010

The wholesale diamond dataset includes numerical data (e.g. price, carats, etc) and categorical data (e.g. clarity, quality of cut, etc.)

There are 407280 total diamonds in the dataset covering sales from 2010 to 2021. For each diamond sale we are given 11 attributes:

1. carat: the diamonds weight. 1 carat = 200 mg
2. cut: rating system from 1 to 5 (poor to ideal)
3. color: standardized color code. Each diamond has a color
4. clarity: standardized table. Measure of any defects that can impact visual appearance.
5. depth: percentation (0 to 100) relating the diamonds depth (top to bottom) with its width
6. table: percentation (0 to 100) relating the diamonds overall width to the width of the top part
7. price: what the diamond sold for
8. x: length in milimeters
9. y: width in milimeters
10. z: height in milimeters
11. year: year of the sale

There are many good internet resources describing these attributes in detail with visuals. You are strongly encouraged to read up on what each of these attributes mean.

Data science eithics: We are excited to set up a project that allows you to explore a robust and real data set with lots of opportunities for learning. We also recognize that the diamond industry has a complicated history including involvement in wars and exploitation practices. If you're interested, we encourage you to learn more about this industry and to consider how the field of data science could be used to support strong ethical practices. For example, is there data you would want to see about diamond sourcing? Is there early data about lab grown diamonds that you would be interested in exploring compared to the data set you have today? All great data work considers ethics in a variety of ways. As you launch your careers in this field, we hope that you do, too.

4 Your tasks as the new data science hire

Before doing any work, please read all of steps 1-4 to get a holistic view of the project. For team work, it is suggested that every member of the team be involved with steps 1 & 2. While all team members should have some involvement with steps 3 & 4, some members could put more focus on building the price prediction model (step 3) while others could focus on the dashboard/app (step 4).

4.1 1. Data cleaning

Check the data for correctness. Remove any data entries that appear problematic. Summarize the data that was corrupted and the problems encountered. An example of a problematic data entry would be a negative sale price, for example. These kinds of data quality problems are commonly found in real-world datasets.

Deliverables: (i) Cleaned dataset as a csv file, (ii) summary of problems with the original dataset (what data was corrupted? How many corrupted entries?).

Tips: Because the dataset is so large, manual inspection isn't a good way to clean the data. Instead, start off with task 2. As you carry out the exploratory analysis, the problematic data entries should become obvious.

4.2 2. Exploratory analysis and summary statistics

Create plots, graphs, and other visuals to summarize interesting aspects of the dataset. This will give you and your company insights into the dataset. Some ideas could include:

- Summary of attributes cut, color, price, etc.
- How many diamonds of each type of cut are there?
- How much do diamonds cost on average? Whats the variance and distribution of prices?
- Generate summary statistics for the attributes.
- How does the diamond cost vary with carat, year, color, and other properties?
- correlations between the variables.
- identify trends
- Clustering
- Use an off-the-shelf algorithm to see if the diamonds in your dataset can be naturally grouped into clusters.

These are just some ideas. You should follow your instincts about the kinds of figures and analyses to perform. For example, a pie or box chart is probably a good choice for showing the population of diamond cuts in the dataset, while a histogram is probably a good choice for showing the distribution of carats.

Deliverables: (i) A summary report (either docx, pdf, a jupyter notebook, a webpage/app, or some other medium) that summarizes the dataset and explain insights you have gained from it, (ii) please put some additional focus on how the attributes of the dataset impact the diamond's sale price. This will be useful for your next task. While you might make any charts for your own personal enjoyment, focus on 7 to 10 top-level charts that provide the best data insights. You can make your summary report part of the website or dashboard built in step 4.

Tips: Start off by thinking about the kinds of figures you want to make. Google around for "exploratory analysis" – there are many excellent blog posts about the approaches you can take for general datasets. Don't get carried away with making tons of figures. At most 10 well-chosen visuals should be enough to give some insights into your dataset.

4.3 3. Building models: price prediction

The main goal of your work is to build a model to predict the price of diamond sales in 2022. That is, your model should take is input the attributes (carat, cut, color, clarity, depth, table, x, y, z, and year) and predict the price (either as a single number or a probabilistic range of values).

Train a machine-learning algorithm to estimate the price of diamonds based on these attributes. There are many off-the-shelf machine learning algorithms for this task. You should try a few and report on their success. Some points to keep in mind:

- What are the most important features for predicting the price? Some attributes, such as carat size, should be strongly correlated with price. Other attributes may only be weakly correlated, if at all. Others could be redundant. Because machine learning algorithms work best with a good "feature set", you should report what features you've tried and what works best (and possibly why!).
- Consider building different models for different clusters/kinds of diamonds. Clusters could be identified with machine learning tools (see step 2) or through intuition. A pricing model for large and small carat diamonds could work better than one model for the entire data set, for example.
- Report on how you trained and validated your model. What was your test-train split and why?
- Report on the different models you tried. How do you assess the accuracy/success of your models? Whats the accuracy of the best one.

Deliverables: (i) A model that predicts the price of a diamond based on the input attributes, (ii) a discussion about the model's accuracy, performance, and limitations, (iii) discussion about how the model is built and what alternative models and modeling approaches you tried.

Tips: The price prediction problem you are solving is known as regression. Some popular machine learning algorithms for regression include

1. polynomial regression
2. Support Vector Regression
3. Nearest Neighbors Regression
4. neural network regression
5. decision tree regressor
6. linear models
7. lasso
8. random forest

… and more. Please keep in mind that your dataset contains a mixture of numerical (carats, price, width, etc) and categorical (cut, color, clarity, etc) data types, and your approach to price prediction should keep this in mind.

Tips: If you are looking for a simple way to get started, pick a year (say 2010) and plot carats vs price. You should notice an obvious trend. You could then build a simple model based only on the relationship between the price and the carats. The model won't be very accurate, but it should give some flavor for how regression models are built and used. You can also use this simple model as you develop code for step 4.

4.4 4. Using your price prediction models

Congrats on building a sophisticated machine learning model to predict prices! Now its time to deploy it. Build a website or dashboard that…

- Summarizes your model: How it was built? What algorithm does it use? Are there any known limitations?
- Allows users of the website to enter diamond attributes and have your website or dashboard report the predicted price
- Allows users to upload a CSV file with diamond attributes (structured the same way as wholesale_diamonds.csv), and your diamond-pricing software should generate a new CSV file with the predicted price.

Now use your diamond pricing software as follows:

- The data file for all diamonds your company plans to sell in 2022 is called **diamonds_for_sale_2022.csv**. Upload this data to your software. Provide a summary report estimating what your model predicts the total diamond sales will be in 2022.

Tips: Depending on the programming language you are using, there are different options for this step. If you are using Python, consider streamlit, Dash, or flask. Jupyter noteboks are discouraged but could be used as a backup. If you are using R, consider shinyapp. Popular dashboards and other tools like Tablaeu, dplyr, and data studio might also work.

4.5 5. Going further

If there's extra time, consider…

- **Data scraping**. Build a data scraping script to get diamond attributes and prices on the internet. Feed this information into your diamond-pricing software to make a "buy" (price is below your estimate) recommendation.
- **Deep networks**: Use deep learning software like tensorflow or pytorch to build a pricing model. Experiment with the number and depth of deep layers, training epochs, and other settings to get a good model. How does this compare with the model you built in step 3?
- **Cut classification**: Your company suspects their newly hired diamond appraiser is incorrectly assigning values of diamond cut. They would like you to devise an algorithm to predict the cut based on other properties, which can then be compared to the appraiser's assignment.

4.6 Appendix: Recommended Python tools

If you are completing this project with Python, consider the following data science tools:

1. Anaconda (https://www.anaconda.com/products/individual): Anaconda will install Python on your laptop and allow you to easily install other Python packages like scikit-learn, pandas, matplotlib, and many other data science tools. There are many excellent online resources for how to setup and use Anaconda.

2. Jupyter notebooks (https://jupyter.org/): Jupyter notebooks allow for interactive coding with a web browser. You can write and run python code interactively right in your browser! This is a great way to explore data, build models, and do other kinds of interactive computing

tasks. Anaconda should automatically install Jupyter notebooks. There are many excellent online resources for how to setup and use Jupyter.

[ ]: