Foobar Inc. Widgets Exercise

Brandon C. Loudermilk

May 3, 2016

Project: Widgets 1.0

All supporting documentation and data: Foobar Inc. Widgets 1.0

WARNING All work should be your own. Working in groups is not allowed. You are competing with your fellow students for a limited number of open positions - **cheating will not be tolerated**. Otherwise you can use any resource available to you including google, stackoverflow, etc. *Don't worry if you can't answer all the questions. You may need to hack out a solution or resort to other tools and techniques to solve these problems.*

If you choose to use SAS, you may use the UI for initial exploration of data, however, this should be complemented by source code including comments. Final output should be a **single printed document** that answers the questions below, providing textual description, code snippets, and data visualizations as needed (if you are familiar with R, think of *R Markdown* .Rmd; if you are familiar with Python, think *iPython Notebook*; otherwise a word document will suffice).

You are a scientist - so you are **expected and encouraged to go beyond the questions** listed below. What additional questions does the data raise? What alternative approaches to missing data might be taken? What techniques might you use to address collinearity? What are the best ways to visually display this information.

Professionalism matters! Are you labels properly labelled? Do your documents have any spelling or gramatical errors? Is your code commented? Are your documents formatted nicely? Etc.

Background: Foobar Inc. is a fictional company in the Bay Area that produces widgets. Widgets have soared in popularity over the past year, and Foobar Inc., has hired you to do an analysis of their data.

Goal: You have been provided a subset of Foobar Inc. data stored in a file called **"widgets.csv"**. Using whatever data science tools (e.g., SAS, R, Python, etc.) and packages (e.g., numpy, e1071, dplyr) you feel comfortable with. You have been asked to do an EDA and develop a model capable of predicting widget price given this historical data.

At a minimum this will entail:

1. reading data
2. munging data
3. producing charts highlighting key relationships
4. determining how to train/test your model
5. training a model(s)
6. testing your trained model
7. improving/refining your model
8. interpreting your results
9. discussing future directions based on your insights/analysis

Questions

For the interview, bring your *commented* source code you wrote as well as any EDA visualizations and written explanations of of your observations and analysis - package as a single well-formatted, professional document.

Specifically you should address the following:

1a) what are the dimensions of the data set?

1b) which variables **should be** numeric?

1c) which variable contains NAs? how many NAs?

1d) what other variable/column might have missing/problematic data?

1e) what is the mean widget price? why might (1d above) be a concern for computing mean *price*?

1f) How will you deal with the missing/problematic data in 1c) and 1d) for purposes of training your data?

2a) create a scatter plot of widget size and price (price on y-axis)

2b) describe the relationshp between size and price

2c) why might you transform these variables? how?

3a) which construction material sold the most units?

3b) which construction material had the greatest mean price?

3c) which combination of construction material and style had the greatest max price?

4a) create a histogram of widget weight

5a) what is the correlation coefficient between widget size and height?

5b) why is this problematic for some models?

5c) how might you deal with this if you want to train a linear model?

6a) create a box plot of prices by widget style

6b) which widget style has the highest median price

7a) Did you split your data? If so, how and why?

7b) How would you ensure that your experimental groups (train/test) are comparable/equivalent?

7c) What would you do if train/test groups are not comparable?

8a) train a linear model on these data - what are the significant variables?

9a) test your model - how did it perform?

10a) how might you increase the performance of your model? *************************************

Data Dictionary for widgets.csv

1.) **widget_id** - *unique widget ID*

2.) **size** - *inflated size of widget cm3*

3.) **construction** - *construction material* {"aluminum", "brass", "bronze", "copper", "nickel", "steel", "titanium"}

4.) **weight** - *weight of widget in grams*

5.) **height** - *height of widget in centimeters*

6.) **zip** - *zip code where widget was sold*

7.) **quality** - *quality of widget* {“terrible”, “bad”, “okay”, “good”, “great”}

8.) **style** - *style/shape of widget* {“circle”, “double-circle”, “double-square”, “double-triangle”, “oval”, “rectangle”, “square”, “triangle”}

9.) **price** - *price of widget ($USD)*