Vandana Anand

MKT 568

Quiz 1 Spring 2020

By answering these questions and emailing me your solutions, you are affirming that you have neither asked for nor received any help from your classmates or other friends. (You may, of course, use any other written or online source you wish.)

Due: March 1, 11:59 pm
Scoring: 31 pts Total
Fill in the blanks where applicable. Otherwise, paste in your answer below the question.

A. Choosing Method (11 pts)

An online real estate company estimates an appropriate asking price for a house based on the actual sales prices and property characteristics of recently sold houses in the same locale. A sample dataset is included in Module 5 of the Canvas site. Some of the characteristics that probably help determine ultimate sales price are house age, house and lot size, number of bedrooms and bathrooms, existence of a garage, pool or sidewalks, and many others. The Neighborhood in which a house is located is also likely to be an important influence. This estimation process is legally very important: some online companies have been sued because of the supposed inaccuracy of their estimates as well as their lack of transparency (that is, the availability and operation of the pricing algorithm).

1. Name three things that might threaten the relevance of these data.
   1. Outliers that may be present that could cause techniques not perform well
   2. Missing data which decrease the sample size and makes the results more variable
   3. Unexpected types of data, such as wrongly classifying numerical, categorical, etc. Too many variables, like dummy variables, can overfit the training data

2. What methods, among those covered in this course so far, can be used?
   We can use our knowledge of the data to develop a context, fill in with the mean or median of the variable, fill in with a random sample from non-missing variable values, and fill in with modeled value, such as regression or k-nearest neighbors.

3. Currently, many realtors set a house's asking price by recent sales of similar houses in the neighborhood, sort of like a nearest neighbor analysis. Name an advantage that a formal nearest neighbor method has over the realtor's informal analysis. Name an advantage of the realtor's method.

   An advantage of a nearest neighbor analysis is that we can use kNN to use the mean of the closest non-missing data in order to predict the missing data. This can decrease the number of gaps and make the results less variable.

   An advantage of the realtor's method is that they do not have to worry about distance from one data point to another so we don't have to apply standardization or min-max normalization before applying the analysis, which makes it cost efficient.

4. If you were fitting a regression model, why might you include interaction terms between Neighborhood and house size?

   I would include interaction terms because there might be similar house sizes in a given neighborhood, so these two variables could have a relationship that could affect the other. We can find out whether there is a relationship between the two variables.

5. One of your team comes to you and says: "My regression model is OK, but I really had higher accuracy from BAGGing a regression tree. Can we use it?" What do you say?

   I would say that Bagging does have its advantages in that it helps reduce variance and avoid overfitting of that data, whereas the regression analysis may be more prone to it. Although bagging does give us a high accuracy, there is a loss of interpretability in the model resulting in a higher bias and it is computationally expensive to perform, which may not be useful in this case when we can use the regression model.

B. Helping Your Team (6 pts)

While analyzing a dataset of house prices as in Question A. above, one of the junior members of your Data Science team comes to you with a technical problem. As in Question A, she is working on a dataset of house prices in a small city (Ames, Iowa) in the Mid-Western US (dataset available on Canvas) and has run into a technical problem with the variable Neighborhood and its role in predicting SalePrice. She says: "I did what you told me and used Neighborhood as an input to a linear regression model of SalePrice. I also divided my dataset into a Training portion and a Test set. My problem is that Neighborhood is a nominal or text variable using the names of areas of this city, and is not a number. And there are 25 Neighborhoods, some with only few observations. So far the model I built on the training set, using Neighborhood, looks different from its prediction on the test set, and I think the problem lies in the Neighborhood variable. Any suggestions?"

You know that SalePrice almost certainly is influenced by where a house is. Name three things your teammate can try to use the idea of a house's neighborhood in her models. (6 pts)

1. You can merge the cities with few observations into another that is similar.
2. You can create a level that contains all small and unimportant levels
3. You can create a new numerical variable, whose values are the average value for each level, preferably from a new dataset, which produces one variable to replace all the dummy variables

C. Extra Findings (8 pts)

Recall that in Module 1, I presented a scenario (the "Story with Morals") in which the company president described my task in the following way: "We think that BAC cable is the most troublesome type we have in our company. We'd like to replace it all, but I only have $30M with which to replace the worst cables. I want you (the Data Miner) to tell me how old the cable should be to warrant its replacement."

I hope you remember that the results of my analysis were surprising, and indicated a different use of that $30M than the president thought. What should you tell the president, and, in particular, how will you explain--in the most persuasive way—why his original request above should be modified? (8 pts)

1. We are prepared to spend our budget of $30 million to replace underground BAC cables because they are vulnerable to wet weather. We want to find out how old these cables will be when we need to replace them. The following items were gathered in order to analyze this question: a sample of various types of cables, such as BAC and others, from one exact year and their ages. This was merged with repair records that stated whether or not a given cable was repaired in that same year. Then, the predictions were graphed using the model with the equation being the function of age for each cable type.
2. Each cable type and its age was used in the model of the probability that a cable segment was repaired.
3. The analysis shows that the BAC cable is getting better with age and does not have a high repair rate. This does not make sense as the more usage of a cable wears it down. So, the conclusion is that the most repair-prone cables have already been replaced and we should efficiently spend the $30 million budget in another area that will save the company money.

What kind of analysis result would you emphasize to an academic audience, but downplay for a non-technical, real world audience? (Note: that result is not necessarily mentioned in the Module 1 slides—it just has to be something that you might cite for school homework.)

For an academic audience, I would talk about the technicalities of the model, challenges in using the model, and compare results from multiple models that I used to arrive at this conclusion such as the accuracy of the predictions from the model. For a non-technical real-world audience, I would overall make the presentation a lot less technical and show only the final results that I obtained by introducing the model, talk about the variables that influenced this prediction and their possible relationship to each other, and explain, in the context of the company, suggestions how to use this model in other areas of the business to improve the approach or methodology.

D. Seeing the Bigger Picture (6 pts)
One of the roles you must play as a data science manager is to take a broader view of any of your team's projects, going beyond the technical application of a technique, and thinking about the overall viability of the project. In this scenario, your client is looking for actions that increase the probability of a customer's churning from a subscription service like cell or video service. From a dataset of last month's subscription information, a classification tree reveals that a customer who receives a refund from his company, has a higher frequency of churning than for the average customer. This higher frequency is nearly 100%. Your client interprets this to mean that using refunds as a customer service tactic backfires by increasing churn.

1. Why is this analysis unlikely to prove the client's contention (be as specific as possible)?
   It is unlikely to prove the client's contention because refund is not a valid predictor of churn since customers get a refund after they churn.

2. What is another interpretation of the relation between refunds and churn?
   Once a customer gets their refund, they have already churned. Due to this, the model predicted that 100% of the customers who get a refund would churn which is true because the model is predicting something that has already happened.

3. Suggest what additional data would be needed to strengthen your speculation in 1).
   We could gather more data, such as a year's worth of data, to see whether a customer will churn in a longer period of time. We could also see a list of the number of their refunds or

the customer’s rating of the product which can indicate their level of satisfaction or dissatisfaction of the product.