Name: **­­­­Shirazi Siddiqui** Date: **18th Feb 2020**

* **Clearly indicate your answer**
* **This is a Quiz - you must work alone!**
* **Open book and notes, and each problem is worth 10points, unless otherwise noted.**
* **Please use the provided excel data set to answer the following questions.**
* **You are to submit the completed Quiz by the start of next class via Canvas**
  + **If late in turning your quiz in, you will automatically lose 30pts**

**Problem 1**

Please carefully take a look at the data set (target = share) provided for training and developing a predictive model. Find all the data issues, and correct the issues (you’ll need it for Problems 2-4). What were the data issues and how did you correct them?

**Issue 1:**



**Issue 2:**



**Issue 3:**



**Issue 4:**



**Issue 5:**



**Solution to Issue 1 :** Effecacy should be from 1 to 10, 100 seems an outlier .We should consider maximum value to populate as 10.

**Solution to Issue 5 :** Price should be from 1 to 10, 15 seems an outlier .We should consider maximum value to populate as 10.

**Solution to Issue 2, 3 and 4** : Empty cell for Safety, Marketing and Effecacy : should be replaced by avg value.

**Problem 2**

Develop a master predictive flow that includes regression, decision tree, and neural networks. This development process is up to you. Please paste the result from the Model Comparison node – show which model is the best one based on the ASE on Validation Data?

**Master predictive flow :**

A close up of a device

Description automatically generated

**Model comparison result:**

**A screenshot of a cell phone

Description automatically generated**

**Multiple Regression Model is the best model as it has least Averaged Squared Error i.e. 0.002245**

**Problem 3**

Please write your stepwise regression model’s mathematical equation here:

**Y = 0.0581+0.0159 (Effecacy) +0.0169 (Safety)**

Where Y is market share

A screenshot of a cell phone

Description automatically generated

P- value of both Effecacy and Safety are significant to the model as it is less than 0.05 level of significance.

**Problem 4**

Develop a Decision Tree model. Please paste your decision tree screen shot here and interpret your decision tree result.

**Default Decision Tree**: It shows Effecacy is the most important variable for predicting Market Share as it has highest log worth i.e. **11.8696**. The default split node is 5.5.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

So, changed Price Node to 4, Safety Node to 7 and Effecacy node to 7 but effecacy is still the most important variable for the prediction of Market Share and log worth for the same is still highest i.e. **6.6708** but price variable seems insignificant to the model as it has “**0.0**” log worth.

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated  
As per Variable importance table, Marketing and Price have “0” importance.

Hence, Marketing and Price are insignificant to the model and below model comparison result shows this interactive decision tree model is better than default decision tree as Averaged Squared error is less than default decision tress.

A screenshot of a social media post

Description automatically generated

**Problem 5 (this is a separate problem – not part of questions 1-4)**

You have built a Neural Nets model using a rich data set. In the summary output, you learn that it has 4 inputs, and 3 hidden nodes. How many total weights does this NN model have? Please show your work.

**4 Inputs :** X1 , X2 , X3 and X4

**3 hidden units :** H1, H2 and H3

**1 Target:** Y

**Yˆ = W00ˆ+ W01ˆ H1 + W02ˆ H2 + W03ˆ H3**

**H1 = tanh(W10ˆ + W11ˆ \* X1 + W12ˆ \* X2 + W13ˆ \* X3 + W14ˆ \* X4)**

**H2 = tanh(W20ˆ + W21ˆ \* X1 + W22ˆ \* X2 + W23ˆ \* X3 +W24ˆ \* X4)**

**H3 = tanh(W30ˆ + W31ˆ \* X1 + W32ˆ \* X2 + W33ˆ \* X3 + W34ˆ \* X4)**

**Answer: Total weights are 19**

**Problem 6**

When you are developing a model, where should you spend most of your time?

1. Developing regression and decision tree models
2. Interpreting the result
3. ***Data preparation***
4. Beautifying the model flow

**Problem 7**

If you are starting a BI function in your company, how would you go about introducing the predictive analytics component of your function so that other business partners can buy in to the results from the predictive models?

On their own, business intelligence solutions are geared and provide insight into what has already happened or is happening right now. With the addition of **predictive analytics**, however, you suddenly become an expert in “***what happens next****”.*

Indeed, there is a wide range of different ways that other business partners can leverage, predictive analytics in business intelligence solutions that are worth exploring.

Let’s explore what they are and see how business can use them to their advantage.

Before you can learn how to better leverage predictive analytics in business intelligence solutions, it's important to come to an understanding about what this concept is actually describing in the first place.

Predictive analytics can be used in a variety of different ways in nearly every industry that can think of. For example:

1. Cybersecurity experts use them to help detect fraud or cyber-attacks while they are still in their early stages, thus giving them a better chance to do something about it before the problem grows too large. Predictive analytics are used to analyze all of the information going on across a network in real-time
2. Other companies use them to optimize their [omni-channel marketing campaigns](https://getcrm.com/blog/omnichannel-marketing/) to reach their own audiences in a more effective way.

What predictive analytics are NOT, however, is a silver bullet that is guaranteed to make an organization better, stronger or faster. As the old saying goes, ***"There is no fate but that which we make for ourselves."*** Predictive analytics are not a guaranteed way to confirm that something is going to happen. Instead, it's a way to give yourself access to better, more informed insights based on what has already happened to better prepare for certain things that are likely to occur down the road.

### Putting Predictive Analytics to Work For You

When it comes to the major nexus between predictive analytics and business intelligence, it's important to keep in mind that your goal is less about predicting the future in a literal sense and is more about putting yourself in a better position to achieve desired outcomes by leveraging past behaviors to your advantage.

Many companies use predictive analytics to gain insights into potential future trends. You see this happen in the electronics industry all the time. Based on market research data about what people are buying, what they're responding to, what they like and what they don't like, you begin to get a better idea of where an entire industry might be headed. This puts certain savvy companies in a position to get in on the ground floor of something poised to be the "next big thing."

In other cases, discovering a potential future trend is just a question of applying what we learned from our trend as it behaves across different seasons and mixing that with the most recent behavior of the trend. Together you can [create a forecast](https://www.dundas.com/learning/webinars/21-02-2018-predictive-analytics-for-everyone) that can take into account the seasonality of your business and your most recent performance so you can better plan ahead.

**Problem 8**

You need to finalize your project data. Please describe your data here, and talk about some evidence of cause-and-effect relationship you found within the data. (Even if you are not sure the data set is what you will use, please describe it here)

The first steps involve collecting data, cleaning, analysis, making some simple plots, drawing basic conclusions, and planning next steps. But data do not give up their secrets easily. While we can use data to understand correlation, the more fundamental understanding of cause and effect requires more. And confusing the two can lead to disastrous results.

I have finalized a “**SALES**” dataset for my project. It was consisting of multiple datasets about employees and their salaries, sales rep details, sales data consisting of net sales amount, COGS, net profit, discounts and product details. As it was the huge dataset, so I have created a pivot table to aggregate data and few sets of plots like line and bar graphs to understand data correctly and able to find hidden patterns so that I can use only required and important fields of the business in order to see cause and effect like product **A** is not doing good as compared to product **C**. When I analyze the data and got to know that product **C’s** discount is quite high as compared to product **A** that’s why customers are more attracted to Product **C** than to Product **A**.

With this dataset I can understand the behavior of each product in generation of sales revenue and how I can improve sales in near future like by reducing discounts and changing sales patterns.

**Problem 9**

Please read: 1) **Chapter 7** in your text book, 2) <https://hbr.org/2014/09/a-predictive-analytics-primer>

1. Yes, I have finished reading Ch7 of the text. **Yes**
2. No, I have not read Ch7. \_\_\_\_\_\_\_\_\_\_

Also, please briefly summarize Harvard Business Review article.

We are using predictive analytics in every organization every day to predict future from past data as nothing has this capability to predict future.

Using predictive analytics, we can measure customer lifetime value(cltv) which tells us how much a customer will buy from the company over a period of time. That’s an analytical prediction of the product or service that customer is most likely to buy next e.g. forecast of next quarter’s sales, used digital marketing models to determine what ad to place on which publisher’s website.

Predictive analytics are gaining in popularity because a manager really need to know in order to interpret results and make better decisions. By understanding basics, you will feel more comfortable working and communicating with others in organization about the results and recommendations from predictive analytics.

The quantitative analysis can be done with past data, statistics and few assumptions.

**The Data:**  Lack of good data is the most common problem for an organization who seeks to employ predictive analytics. To make predictions about what customers will buy in the future, you need to have good data on who they are, what they have bought in the past, the attributes of those products and some demographic attributes of the customer (age, gender, residential location etc.).

**The Statistics:** An analyst hypothesizes that a set of independent variables Like gender, income etc. are statistically correlated with the purchase of a product for a sample of customers. The analyst performs a regression analysis to see just how each variable are correlated with each other. This usually requires some iteration to find the right combination of variables and the best model. It’s quite likely that the high scoring customers will want to buy the product, assuming the analyst did the statistical work well and that the data were of good quality.

**The Assumptions:**  The big assumption in predictive analytics is that the future will continue to be like the past. As Charles Duhigg describes in his book “**The power of habit**”, people establish strong patterns of behavior that they usually keep up over time. Sometimes, however, they change those behaviors, and the models that were used to predict them may no longer be valid. What makes assumptions invalid? The most common reason is time. If your model was created several years ago, it may no longer accurately predict current behavior. The greater the elapsed time, the more likely customer behavior has changed.

With these fundamentals in mind, here are a few good questions to ask your analysts:

* Source of data used in analysis.
* Sample data are representative of the population?
* Any outliers in data distribution? How did they affect the results?
* What assumptions are behind analysis?
* Any conditions that would make assumptions invalid?

Even with those cautions, it’s still pretty amazing that we can use analytics to predict the future. What we have to do, is to gather the good and correct data, do the right type of statistical model, and be careful with what assumptions we are making.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | |  |  | |  | |
|  | | | | | | | |
|  |  | |  |  | |  | |
|  |  | |  |  | |  | |
| **Problem 10**  Copy the below data set to Excel, and show how you would you organize this survey data for a food product for predictive purposes? | | | | | | | |
|  |  | |  | |  | |
| **ID** | **Attribute Type** | | **Attribute Score** | | **Market Share** | |
| 231 | Price | | 7 | | 33% | |
| 231 | Packaging | | 5 | | 33% | |
| 231 | Taste | | 5 | | 33% | |
| 231 | Nutrition | | 8 | | 33% | |
| 232 | Price | | 6 | | 28% | |
| 232 | Packaging | | 5 | | 28% | |
| 232 | Taste | | 5 | | 28% | |
| 232 | Nutrition | | 7 | | 28% | |
| 234 | Price | | 3 | | 15% | |
| 234 | Packaging | | 5 | | 15% | |
| 234 | Taste | | 5 | | 15% | |
| 234 | Nutrition | | 4 | | 15% | |
| 235 | Price | | 5 | | 18% | |
| 235 | Packaging | | 5 | | 18% | |
| 235 | Taste | | 5 | | 18% | |
| 235 | Nutrition | | 5 | | 18% | |
| 244 | Price | | 5 | | 35% | |
| 244 | Packaging | | 5 | | 35% | |
| 244 | Taste | | 4 | | 35% | |
| 244 | Nutrition | | 9 | | 35% | |
| 246 | Price | | 6 | | 22% | |
| 246 | Packaging | | 5 | | 22% | |
| 246 | Taste | | 7 | | 22% | |
| 246 | Nutrition | | 8 | | 22% | |
| **Data should be reorganized as below** | |  |  | |  | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Price** | **Packaging** | **Taste** | **Nutrition** | **Market Share** |
| 231 | 7 | 5 | 5 | 8 | 33% |
| 232 | 6 | 5 | 5 | 7 | 28% |
| 234 | 3 | 5 | 5 | 4 | 15% |
| 235 | 5 | 5 | 5 | 5 | 18% |
| 244 | 5 | 5 | 4 | 9 | 35% |
| 246 | 6 | 5 | 7 | 8 | 22% |