STA 380 - Exercises 1

Chase Slocum

August 5, 2016

Probability Practice

Part A

Given the law of total probability $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$, the probability of yes among truthful answers must be:

$$P(Yes) = P(Yes|Truthful) * P(Truthful) + P(Yes|Random) * P(Random)$$

Knowns:
$P(Yes) = .65$
$P(Truthful) = .7$
$P(Yes|Random) = .5$
$P(Random) = .3$

Unknown:
$P(Yes|Truthful) = ?$

$$.65 = P(Yes|Truthful) * .6 + .5 * .3$$

$$P(Yes|Truthful) = .71$$

Part B

Given Bayes' Rule, the probability of having the disease if someone tests positive is:

$$P(Disease|Positive) = \frac{P(Disease) * P(Positive|Disease)}{P(Positive)}$$

Knowns:
$P(Positive|Disease) = .993$
$P(Negative|Healthy) = .9999$
$P(Positive|Healthy) = .0001$
$P(Disease) = .000025$
$P(Healthy) = .999975$
$P(Positive) = P(Positive|Healthy) * P(Healthy) + P(Positive|Disease) * P(Disease)$

Unknown:
$P(Disease|Positive) = ?$

$$P(Disease|Positive) = \frac{P(Disease) * P(Positive|Disease)}{P(Positive|Healthy) * P(Healthy) + P(Positive|Disease) * P(Disease)}$$

$$P(Disease|Positive) = \frac{.000025 * .993}{.0001 * .99975 + .993 * .000025}$$

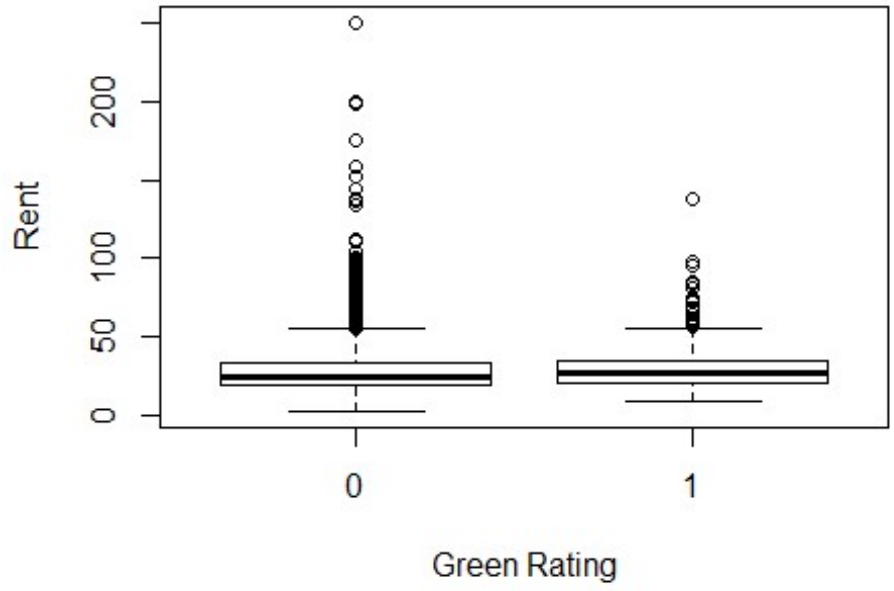
$$P(Disease|Positive) = .1989183$$

Given that the probability of having the disease when someone tests positive is less than 20%, the implementation of this test would lead to more false positives than true positives and would cause a lot of unecessary worrying in addition to money and time lost confirming a positive result.

Exploratory Analysis: green buildings

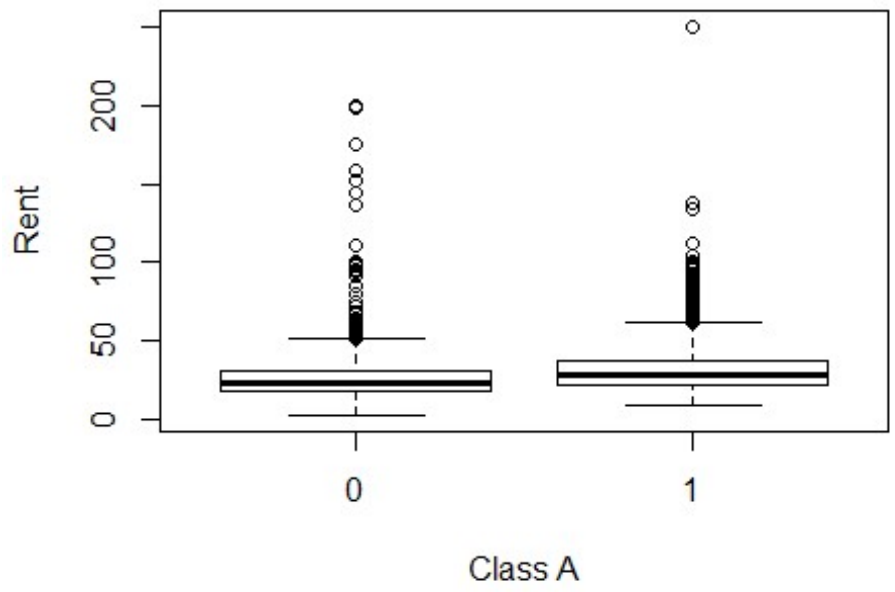
The stats guru's conclusion is far from definitive. The guru's intuition to look at the median as a measure that is more robust against outliers is correct. However, narrowing the decision point to the median ignores a variety of factors.

First of all, simply looking at a boxplot reveals that there is significant variability in the rent per square foot for both green and non-green certified buildings.



Clearly, there is significant overlap in the range and distributions of rent for both types of buildings. This alone raises serious doubts about the validity of the guru's conclusion.

There is also the issue of confounding factors. By just considering medians, the guru has ignored the possibility that factors that are correlated with a green rating are driving the premium in rent. One such factor is the indicator for class A building quality. Intuitively, higher quality buildings will be able to charge more for rent. Looking at a boxplot of class_a versus rent, class A buildings tend to charge more for rent, albeit by a very slight amount.



Following the guru's example to look at medians, the median for class A buildings is 28.2 while the median for lower quality buildings is 23.5.

How does this relate to green ratings? Green rated buildings tend to be class A buildings:

```
       green_rating
class_a    0    1
      0 4598  139
      1 2611  546
```

In fact, 0.7970803 of all green buildings are class A buildings. The median price for green class A buildings is 28.44 while the median price for non-green class A buildings is 28.2. The benefit of going green under the guru's criteria is clearly lost among class A buildings. Among non-class A buildings, the median rent for green buildings is 25.55 and for non-green buildings it is 23.43, so if the building is not class A, there might be potential for the monetary benefits to be present. Similar relationships exist for other variables, including amenities.

If the guru wishes to improve their predictions, they need to revisit their analysis and attempt to account for these other confounding variables, The guru then needs to indentify

where the developer's building falls among the different criteria to determine if going green is worth the investment.

Bootstrapping

The following chart displays the results of five different Monte Carlo simulations. Each one took an individual ETF and simulated holding it for 20 days with a $100,000 initial investment. The chart details the average return and the amounts for which the investor stands a 5% chance of gaining/losing that much or more.

Ticker	Average	5% chance	5% chance
SPY	1020.6846383	-6216.1706941	8347.6744992
TLT	888.4728507	-6068.9610458	8136.7492097
LQD	519.6685899	-2118.6118107	3111.3478299
EEM	200.6093151	-1.039182310^{4}	1.102506710^{4}
VNQ	1313.5357888	-7835.0997183	1.054612510^{4}

Clearly, the safer ETF's are the treasury bonds and the coporate bonds. The domestic equities are a moderate risk/reward fund with EEM and VNQ being riskier investments.

Consequently, the safe portfolio will comprise 40% in LQD, 40% in TLT, and 20% in SPY. The aggressive portfolio will comprise 40% in EEM, 40% in VNQ, and 20% in SPY.

The following chart shows the value-at-risk for each portfolio after a 4-week trading period:

Portfolio	Value-at-risk
Even	-3868.6233562
Safe	-2524.3017099
Aggressive	-7645.1446235

This is a breakdown of the returns of each portfolio that can be used to determine the best investment strategy. It includes the average return, standard deviation of returns, and 90% confidence interval for returns.

Portfolio	Average	standard Dev.	5% Loss	5% Gain
Even	787.1220959	2830.070399	-3868.6233562	5436.244069
Safe	766.8382506	2007.1399657	-2524.3017099	3997.9952506
Aggressive	807.6147542	5230.0099161	-7645.1446235	9438.5953962

Market Segmentation

Via kmeans++, I was able to indentify 6 unique customer segments. 6 clusters turned out to be a good balance between shrinking errors and creating distinct classifiable clusters. These are the cluster sizes:

```
[1] 1349  886 1389  256 1962 2040
```

These are the top words for each cluster:

```
      [,1]               [,2]             [,3]
 [1,] "health_nutrition" "tv_film"        "fashion"
 [2,] "personal_fitness" "art"            "cooking"
 [3,] "outdoors"         "music"          "beauty"
 [4,] "cooking"          "uncategorized"  "online_gaming"
 [5,] "food"             "small_business" "college_uni"
 [6,] "eco"              "college_uni"    "sports_playing"
 [7,] "dating"           "crafts"         "photo_sharing"
 [8,] "spam"             "current_events" "dating"
 [9,] "home_and_garden"  "home_and_garden" "music"
[10,] "uncategorized"    "travel"         "uncategorized"
      [,4]              [,5]             [,6]
 [1,] "adult"           "chatter"        "sports_fandom"
 [2,] "spam"            "shopping"       "religion"
 [3,] "eco"             "photo_sharing"  "parenting"
 [4,] "small_business"  "current_events" "news"
 [5,] "outdoors"        "eco"            "politics"
 [6,] "uncategorized"   "business"       "food"
 [7,] "computers"       "home_and_garden" "school"
 [8,] "home_and_garden" "uncategorized"  "automotive"
 [9,] "travel"          "small_business" "family"
[10,] "parenting"       "music"          "computers"
```

It is clear that most of NutrientH2O's customers demonstrate interests in being active and engaged. The larger clusters (1,3,5,6) have many interests associated with being informed/up-to-date (e.g. current events, politics, fashion). Looking at individual clusters, it appears 1 is related to people interested in the outdoors and being active. In contrast, cluster 2 seems to be populated by more artisitically driven customers, but it is also one of the smaller clusters. The larger clusters (5 and 6) are distinct from each other. Whle both seem to have interests in keeping up with current events or the news, cluster 5 seems more driven by a customer that is a social media user as demonstrated by the prevalence of "chatter" and "photo sharing". On the other hand, cluster 6 seems to be more related to parents interested in investing time and money in their family as evidenced by "religion", "food", "automotive", and "family".