Grace Cheung

gvc8

CPSC 310

HW#2 Part A - Report

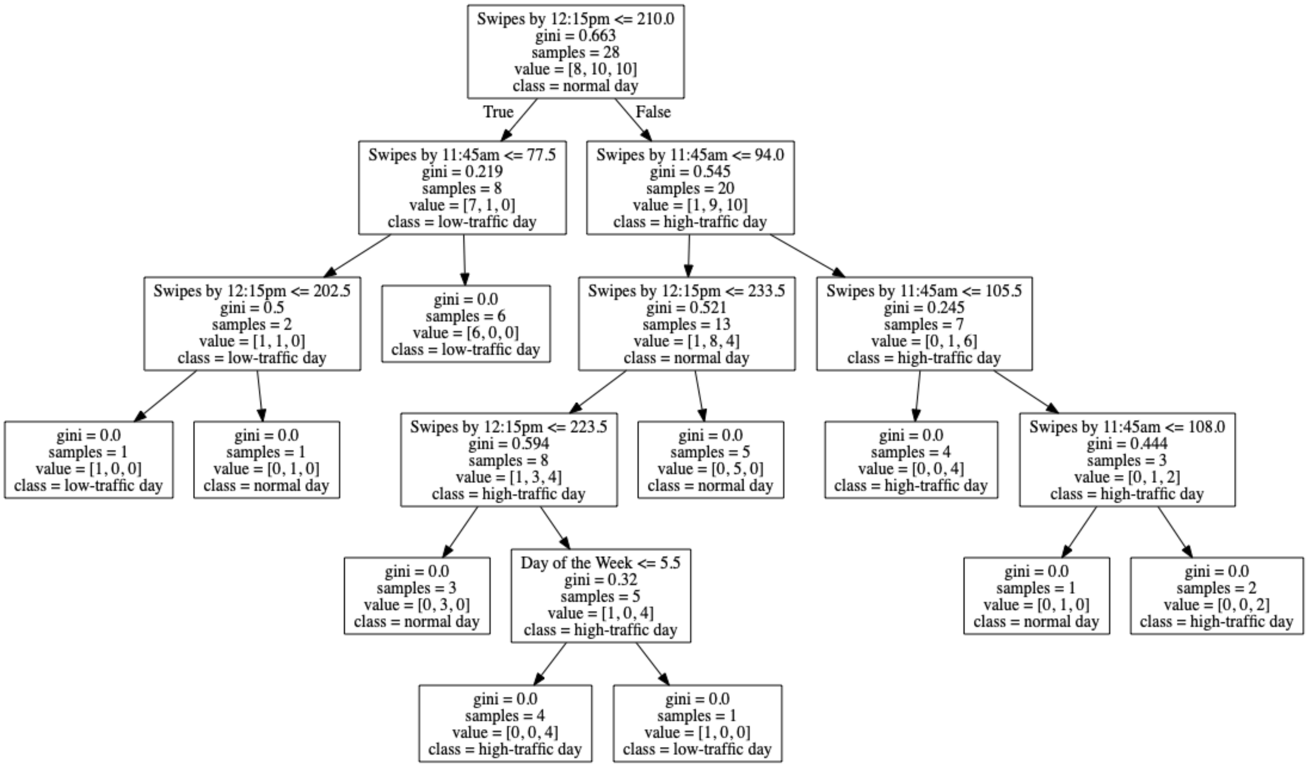
2/21/2020

Part A

I created these decision tree representations by exporting the decision tree as a .dot file using graphviz, and then exporting them as .png files with this code in the terminal:

```
dot -Tpng 10.dot -o tree_BR.png
dot -Tpng 62.dot -o tree_SM.png
```

Branford prediction (tree_BR.png):

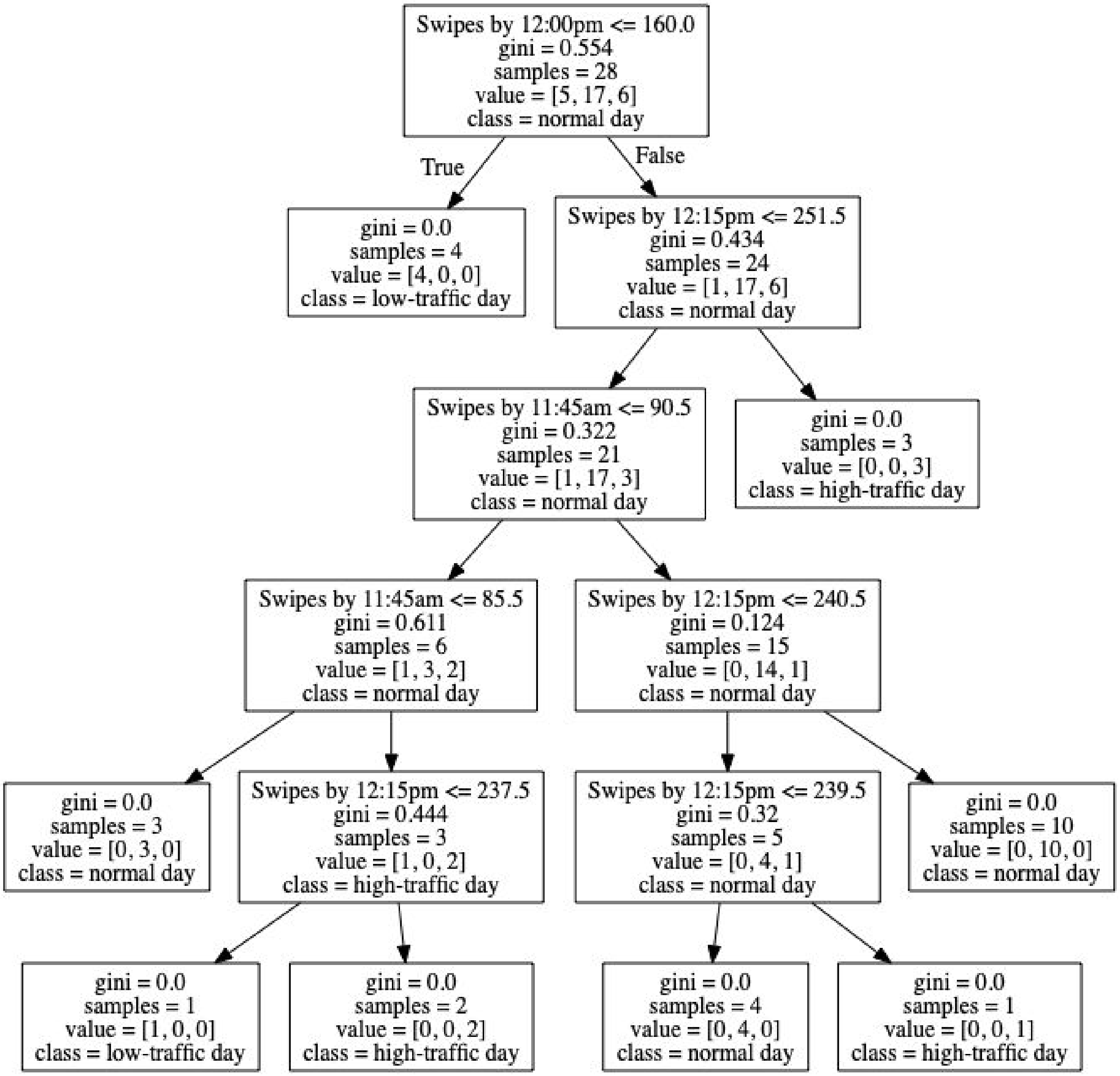


The decision was that it would be a high-traffic day. The root node checks the number of swipes by 12:15pm at a threshold of less than or equal to 210.0 swipes. In this situation, there are 232 swipes, so it is false. So, the first split moves to the right. This checks if the number of swipes by 11:45am are less than or equal to 94.0 swipes. There are 78 swipes, so this is true. The number of swipes by 12:15pm is 232, which is less than 233.5, so it moves to check if it is less than 223.5 swipes, which is false. Then, it checks if the day of the week is less than 5.5. It is a Monday (1), so it is true, giving a high-traffic day end node.

To analyze this decision tree classification, based on total gini values, swipes by 12:15pm and then swipes by 11:45am seem to have the highest contribution to the final prediction. The first split is highly important, as if swipes by 12:15pm are fewer than 210, it almost basically means it will be a low-traffic day, and if greater, then it will basically almost always be normal or high-traffic. Along the same lines, the second split (if we move to the right) seems to basically determine that it will be high-traffic. In this situation, we are traveling the middle path. This third and fourth split are interesting because they imply that between 223.5 and 233.5 swipes by 12:15, day of the week plays a key role in predicting traffic. If it is a weekday, then swipes in between this range predict high-traffic. But, if it is a weekend, then it predicts low-traffic.

Generally, this decision tree makes sense to me. It makes sense that the biggest contributors would be swipes by 12:15pm and then 11:45am, because the two numbers together give a general growth trajectory of lunch traffic for the day, and swipes by 12:00pm would be pretty much factored into the 12:15pm figure. It also makes sense that day of the week would be a later contributor in the decision tree, but it is shocking that it is such a tight range (10 swipes), and that the weekend vs. weekday can make such a huge difference in predicted traffic. I think this is an interesting section, because it implies that a swipe number of, for example, 215, by 12:15pm would predict a normal day for both weekends and weekdays, while a swipe number of 225 on a weekend would predict a low-traffic day. I hypothesize that this could make some logical sense, because the range of times in which people swipe in on the weekends for brunch is much wider. Brunch also starts earlier. So, a swipe number of 225 by 12:15pm on a weekend could mean that the rate of traffic is actually more slow than the same number for a weekday. To explain the strange implication, perhaps the decision tree found a pattern in which overall traffic slowed way down after 12:15pm on weekends if many people swiped in before 12:15pm. This could make sense because most people eat brunch in their own college, which means a cap on the total number of swipes, whereas many people with classes near Branford eat there for lunch during the week. This is still somewhat of a stretch, in my mind though, and is the part of the decision tree that I have the most disagreement with.

Silliman prediction (tree_SM.png):



The decision was that it would be a low-traffic day. The first split checks the number of swipes by 12:00pm at a threshold of 160.0. In this situation, there are 171, so it is false which moves us to the right side of the tree. Then, it checks the number of swipes by 12:15pm at a threshold of 251.5. There are 230, so this is true. Then, it checks for swipes by 11:45am at a threshold of 90.5. It is 90, so this is true. Then it checks the same feature (11:45am) again, but at a lower threshold of 85.5 which is false. Then, it checks for swipes by 12:15pm at a threshold of 237.5, which is 230 and true, predicting a low-traffic day.

The most significant feature of this decision tree is that the possibility of a low-traffic day is almost completely ruled out by the first split, but then the gini value of the fourth split (swipes by 11:45am <= 85.5) is so large, and makes such a huge contribution, that it leads to a low-traffic day prediction.

I think it is interesting that in comparison to Branford, there is less clear path of calculation. It makes sense that day of the week does not factor as much into this decision tree as it does for Branford. Empirically, more people meet friends for brunch in Silliman or have group brunches there. In general, Silliman is more often used by people from all different colleges, while Branford is only used by Branford students and for occasional location convenience. So, Silliman would just have a much tighter range of traffic. This seems to be reflected in this decision tree. Rather than having much clearer splits that demonstrate possible trends, this decision tree seems to be attempting to calculate traffic based on pure numbers. This is especially evident when using the 12:15pm feature. It appears in many nodes, with very similar thresholds across the tree. In general, it makes less sense than the Branford one, indicating that maybe Silliman's lunch traffic does not have as clean-cut of a pattern, and is more variable but with a narrower range.