Hardik Shahu

Customer Brand Preferences Report

Overview of the Task

Find a way to accurately predict the missing data for customer's brand preference.

Gathering and Preparing the Data

For this project, we were given survey results as a couple of CSV files. the attributes of the data were salary, age, education level, primary car, zip code, credit available and their brand preference between Sony and Acer.

Unfortunately, a portion of the data was not recorded correctly which led to brand preference being incorrect for some entries. As such, the data that was provided to me was split into two files, one having the complete and accurate results meanwhile the second file had incomplete results.

In terms of data preparation, there was not much needed as the only modification we had to do was to convert brand column from integer type to category type as we will be doing classification and in order to do that, or data must be categorical.

Looking at the Data

This is a small snipped of the data:

salary	age	elevel	car	zipcode	credit	brand
119806.5	45	0	14	4	442037.7	0
106880.5	63	1	11	6	45007.18	1
78020.75	23	0	15	2	48795.32	0
63689.94	51	3	6	5	40888.88	1
50873.62	20	3	14	4	352951.5	0
130812.7	56	3	14	3	135943	1
136459.3	24	4	8	5	80500.56	1
103866.9	62	3	3	0	359803.9	1
72298.8	29	4	17	0	276298.7	0
37803.33	41	1	5	4	493219.3	1
63704.26	48	4	16	5	299460.2	1
128999.9	52	1	6	0	152232.5	0
57991.3	52	3	20	4	227743.7	1
82474.58	33	4	13	3	424657.5	0
63988.97	62	2	6	3	262136.3	0
132310.4	23	1	11	4	451927.6	1
113236.4	24	2	7	7	198381.9	1
120525.6	57	1	6	2	15964.68	0
121457.6	50	2	2	8	415359.2	0

How do we predict the missing Data?

In order to do this, we will have to build machine learning algorithm models.

Then we will compare how multiple models fair against each other and pick the one that is the most accurate.

From then on, we will be able to use that model to accurately predict the missing data. (Assuming that we find a model that is satisfactory for our case)

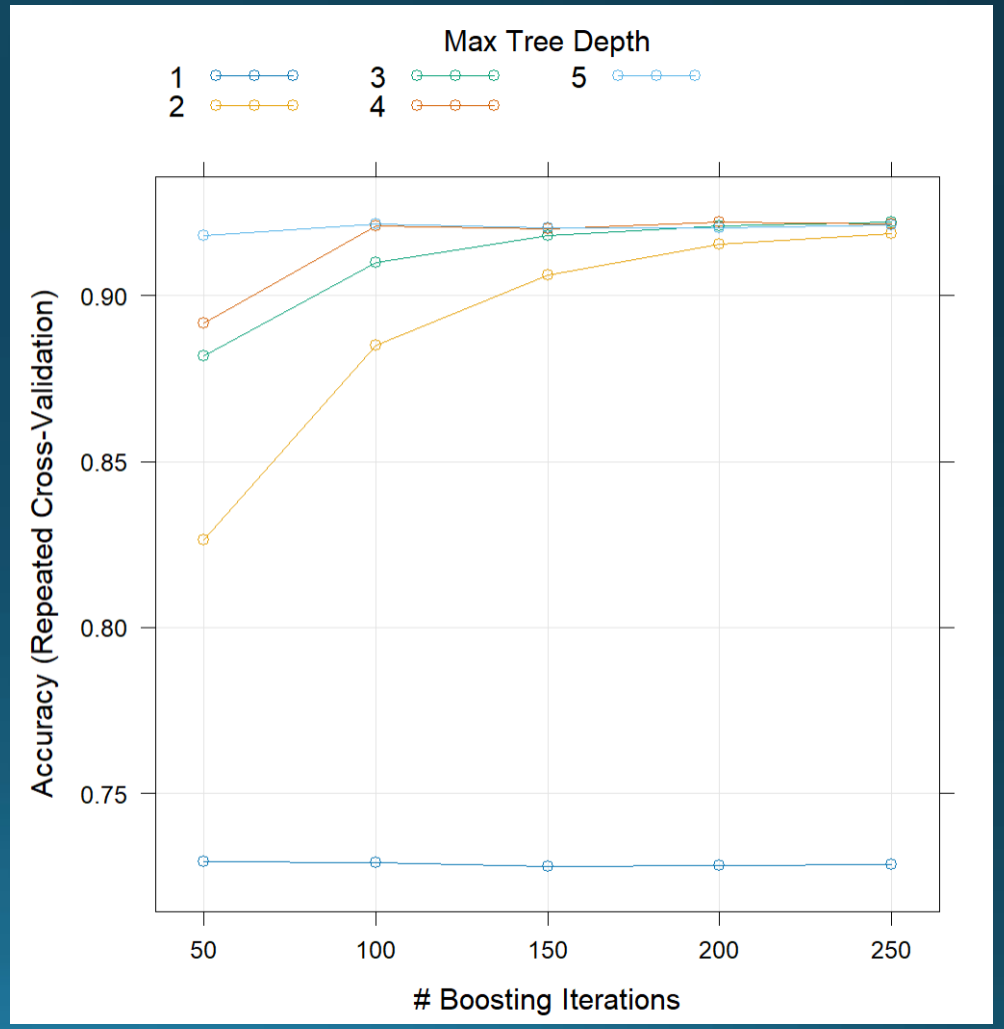
Model 1: GBM

The first model we tried was named Gradient Boosting Machine, or GBM for short. Which combines multiple decision tree's predictions to make a final prediction.

Model 1: GBM (Cont.)

For this model we tried with five different tree depths and iterations ranging up to 250.

Here is how each of these faired as shown on the figure to the right.

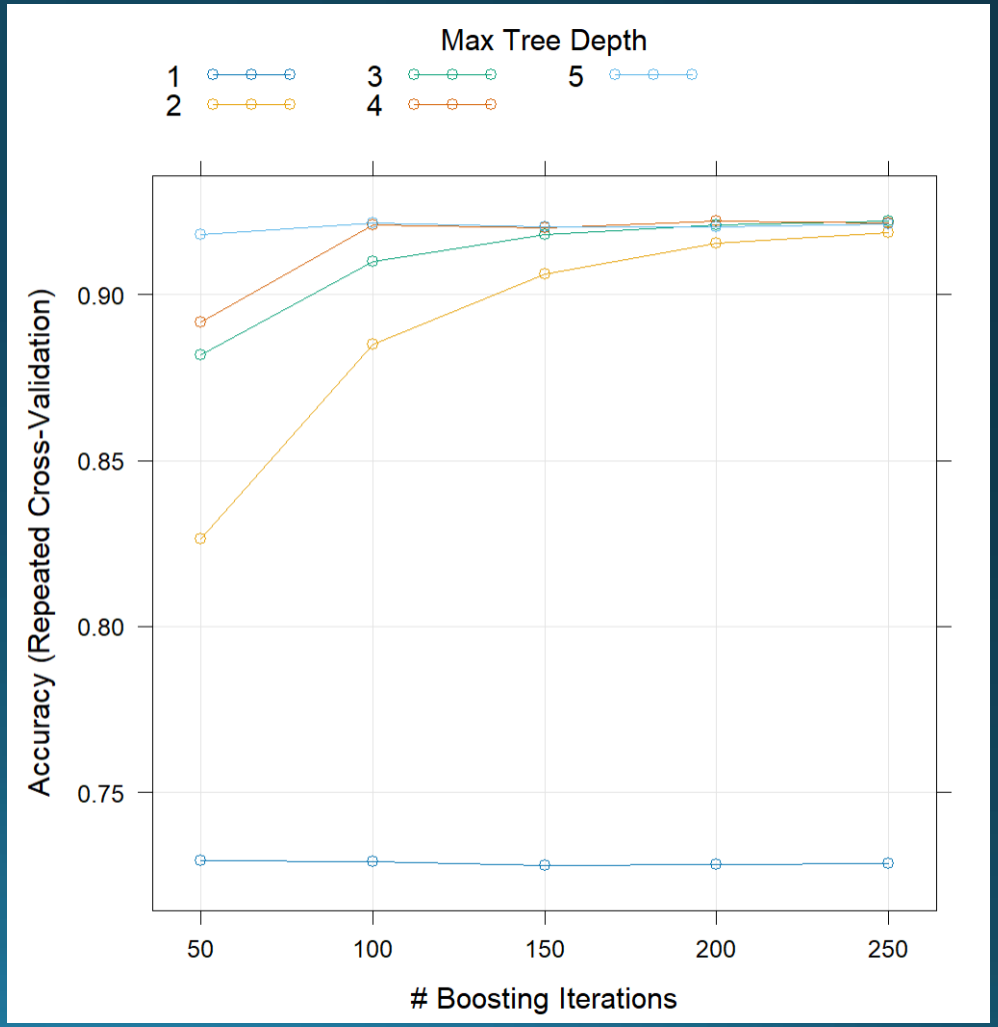


Model 1: GBM (Cont. 2)

As you can see, if we use 1 or 2 max tree depths and stick to the 50 iterations, our accuracy is relatively low compared to if we use 3-5 tree depths. (Accuracy here is determined from a scale 0-1 where closer to 1 means more accurate)

Increasing the iterations to 250 gives us effectively the same results (about 0.92) for tree depths of 2-5 and this is very accurate.

If we were to pick this model, we should use a max tree depth of 5 as it gives the highest accuracy with the least number of boosting iterations.



Model 2: RF

The second model we tried was named Random Forest, or RF for short. Which , like GBM, also combines multiple decision tree's predictions to make a final prediction. However, it doesn't work exactly like GBM as its results end up being a bit different than GBM's and it also varies in how fast it finishes.

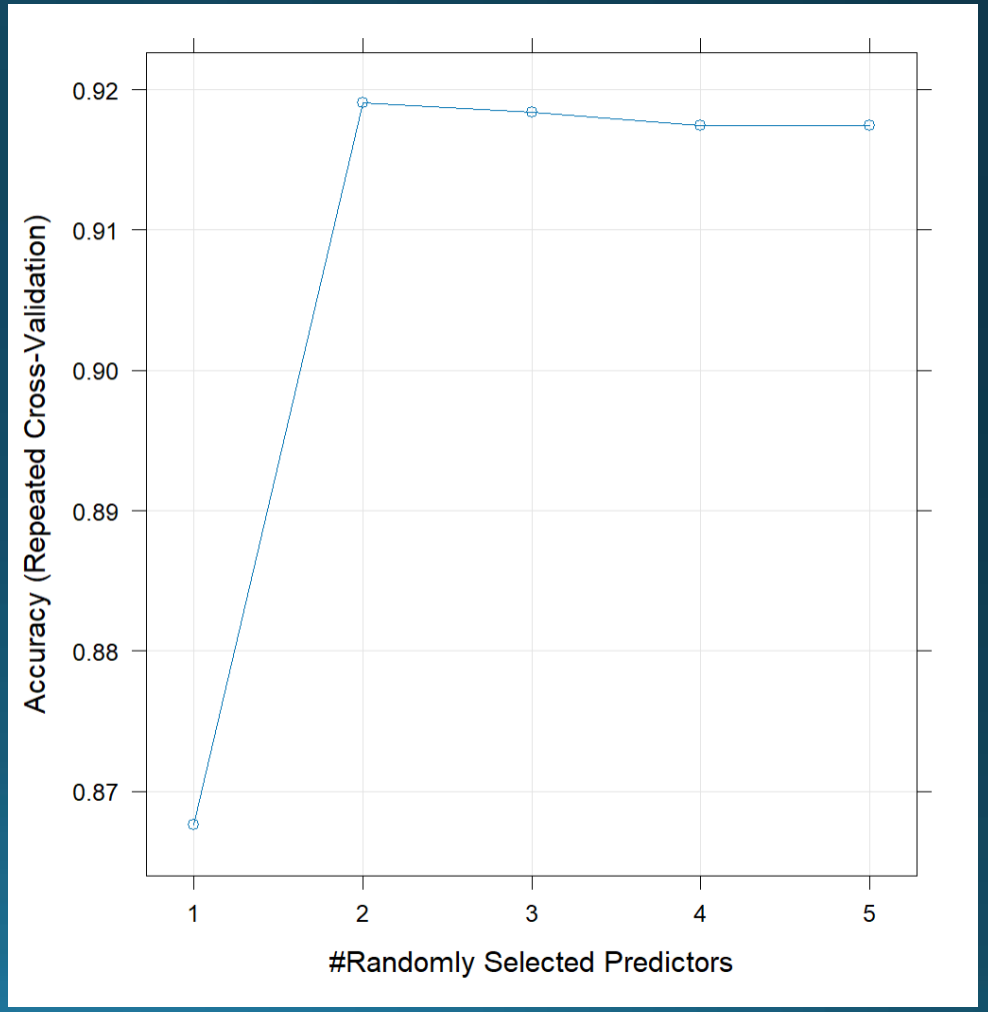
Model 2: RF (Cont.)

For this model we tried with predictors ranging from 1-5.

The figure on the right shows how each of these fair against one another.

As you can see, 1 is worst and surprisingly 2 is the highest (.919) as if we use 3, 4 or 5, we see a slight dip in accuracy.

So, if we were to pick this algorithm, we will should use 2 predictors as it seems to give the most accuracy.



Which Model do we pick?

- Both model's highest accuracy is around 0.92, which tells us that both are excellent choices to build a model with as their accuracies are very high.
- This means we are free to choose either one of these algorithms to use to predict to missing data.
- For now, we will choose the GBM model but once again, we could have gone with either.

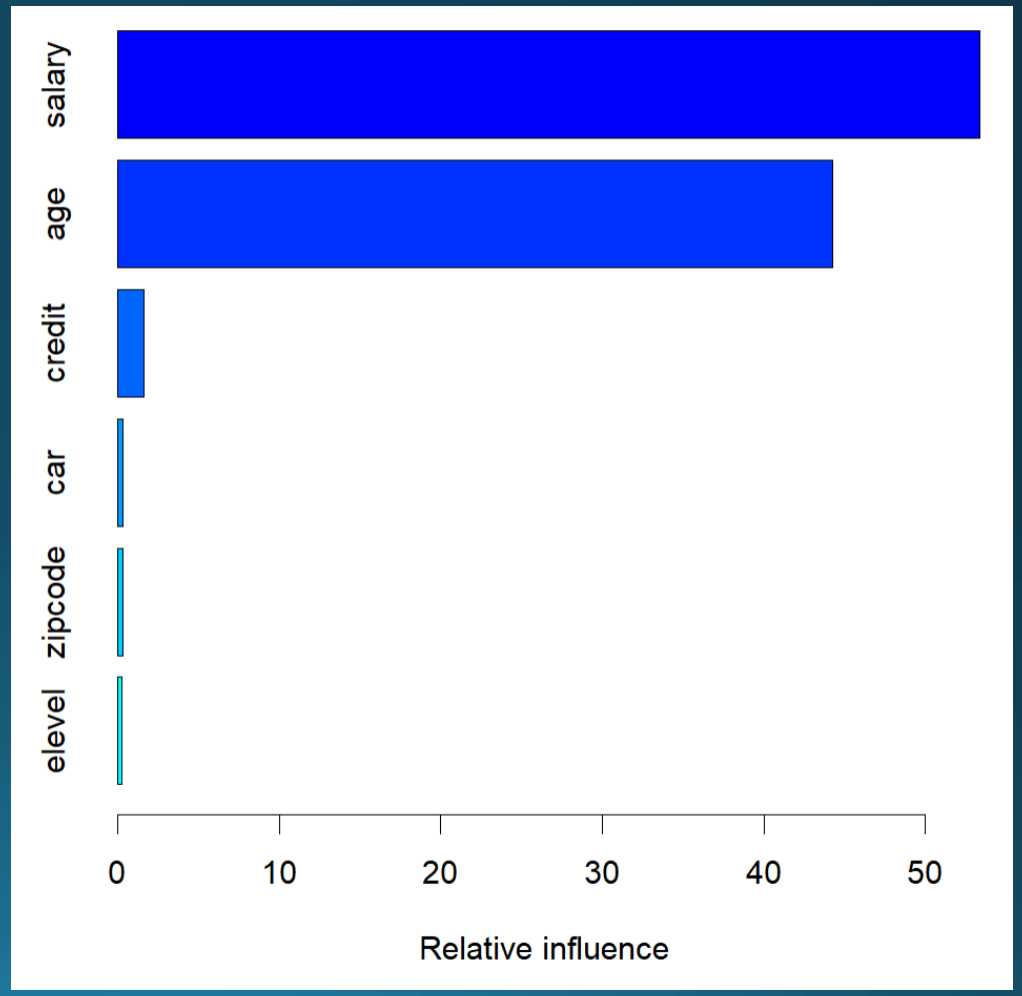
Looking a bit deeper into GBM

Now we will see which attributes were the most influential for GBM to predict.

The figure on the right shows how influential each of them relative to one another.

As you can see, age and salary were the highest while the rest had little to no influence.

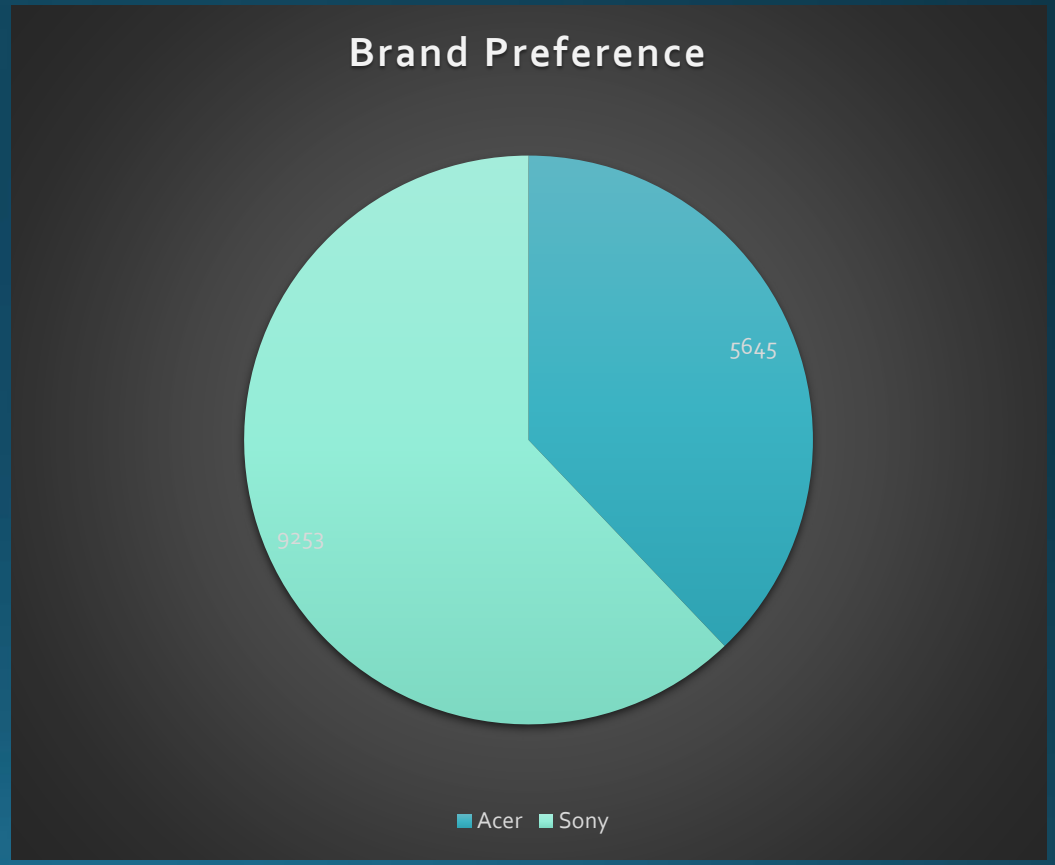
In order words, if we only wanted to predict brand preference, then we only needed to ask them age and salary as a question, but it will not hurt to ask other questions as if we ever want more info or want to predict other attributes, we will have an easier time.



What Brand do Customers prefer?

The pie chart on the right shows the customer's brand preference. This is for all the data, adding the completed results and prediction of the incomplete results.

Overall, it seems most a solid 2/3rd prefer Sony over Acer.



Conclusion

The original task was to find a model that can predict customer's brand preference and we were to do that with the GBM model.

Additionally, we found a bit more insight as when we had all the data, we saw that most customers prefer Sony over Acer, but not all as 1/3rd still chose Acer.