SPRINGBOARD DATA SCIENCE CAREER TRACK

STARBUCKS CAPSTONE (PREDICT NEXT PURCHASE TIME) FINAL REPORT

Introduction

We always ask our customers to come again. After all, repeat customers form the core of our business. Providing fast, friendly service and asking for future business goes a long way, but we need to do more than that. We need to find out when our customers will return.

No, we don't have to ask our customers when they'll be back, we can learn that ourselves by predicting the time of their next purchase. This way, we can make the best use of our incentive offers. If a customer is predicted to return within two days, there's no need to send that customer a discount or a buy-one-get-one offer during those two days. If, however, that customer hasn't made another purchase after two days, then we can rein that customer back in with an offer.

Our machine learning model will help keep those regulars coming back consistently by predicting a time frame within which they'll make their next purchase.

Methodology

Provided features

We have data on 9,130 customers over a 30-day timeline. The time of each purchase is indicated by the hour during that time frame (from 0 to 714). The data contains the value of each purchase as well as demographic information for each customer (age, gender, income, how long the customer has been a member on the app).

Since this model essentially is trying to predict the future, we used the last 10 days of the timeline as the "future." We truncated our timeline at 20 days and the only data used from the last 10 days was the earliest purchase the customer made. We then subtracted the last purchase from the first 20 days from that to get our "next_purchase" variable.

For example, If a customer made a purchase at hour 502 (last 10 days), and that customer's latest purchase in the first 20 days came at hour 442, then we get 60 for "next_purchase," meaning that customer's next purchase came 60 hours after his or her previous one.

Derived features

Using the features we were provided, we derived more granular features. Customers were placed into four groups, or clusters, in each of the following categories:

Recency: How long before the 20-day cutoff was the customer's latest purchase?

Frequency: How often did the customer purchase within the first 20 days?

Revenue: How much money did the customer spend in the first 20 days?

We also segmented the customers by High-Value, Mid-Value and Low-Value based on these clusters.

Finally, we dropped all customers with less than three purchases within the first 20 days, and we calculated the mean and standard deviation of the time differences between each remaining customer's last three purchases.

This process gave us 21 variables to help predict our target variable.

Target variable

Our goal was to predict which one of three classes each customer would fall into.

- Next purchase within 48 hours (Class 2)
- Next purchase within 48-120 hours (Class 1)
- Next purchase in more than 120 hours (Class 0)

Originally, the middle class was 48-96 hours, but that proved too challenging for our model to predict as we will demonstrate.

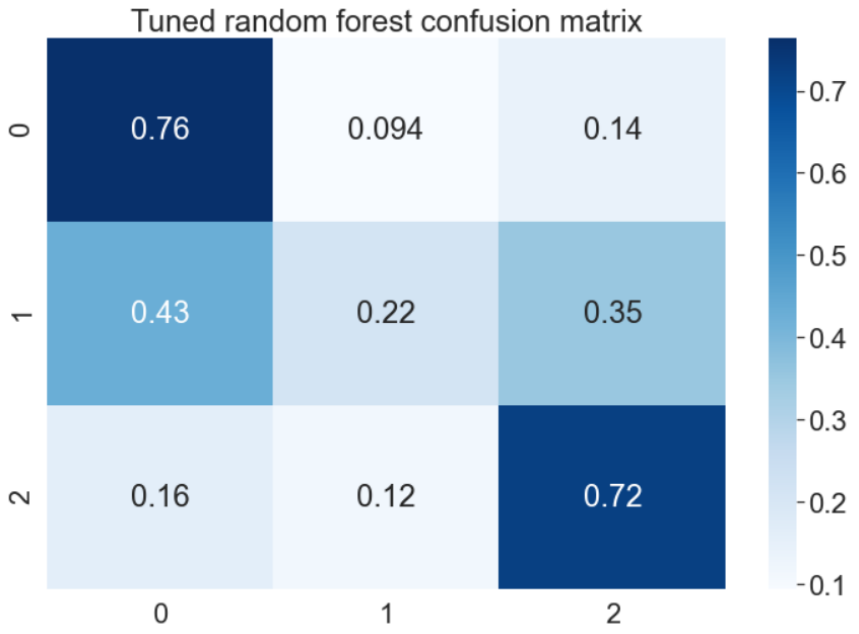
Modeling

Let's start by showing the best model when Class 1 was 48-96 hours. This was a Random Forest model with 61 percent accuracy.

Below is a confusion matrix. The horizontal rows represent the true class and the vertical columns represent the predicted class. So the top left square, the center square and the bottom right square represent the accurate predictions.

The Class 1 predictions, then, are represented in the middle row going across horizontally. We see that only 22 percent of our Class 1 customers are predicted accurately. This is a class that we want to get right because these customers are expected to return in 2-4 days, and this model predicts that 43 percent of them won't come back in at least 4 days if they

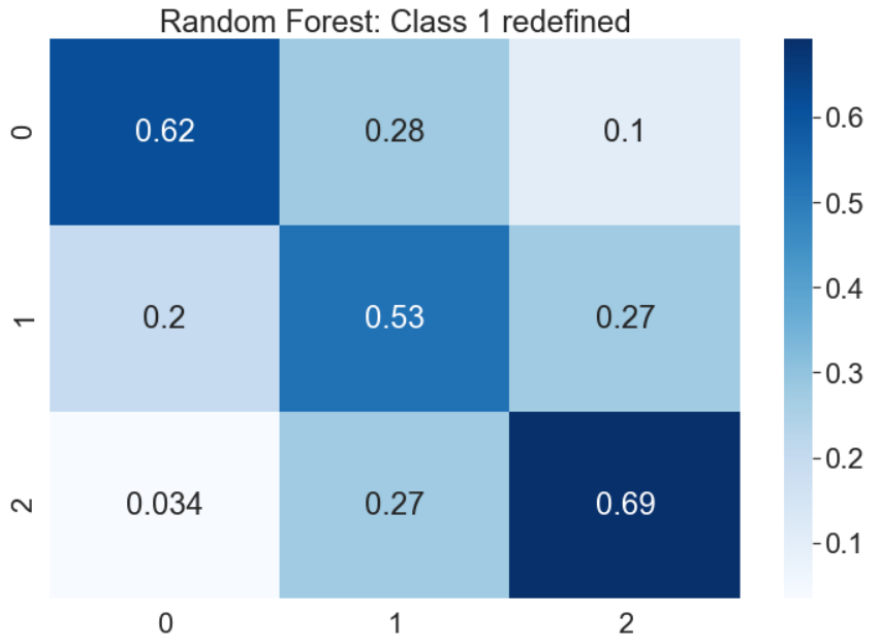
come back at all. That means that we're missing a lot of customers when we send out offers.



Solution

We changed our definition of Class 1 to 48-120 hours, meaning that these customers would be predicted to make their next purchase in 2-5 days.

Now, 53 percent of our Class 1 customers are predicted accurately and only 20 percent are predicted as Class 0 (coming back in 5 or more days if they come back at all).



Hold on a minute here

While we focused on getting Class 1 right, we didn't lose sight of the fact that it's even more important to get Class 2 right. Those are our regulars, the ones expected to make their next purchase within 48 hours.

We see that we predicted 72 percent of them accurately with the first model but only 69 percent with the second one. Sometimes there's a give-and-take with model predictions.

We also considered the errors we most wanted to avoid, ranked in order.

Class 2 predicted as Class 0 (bottom left on matrix): Customers who will come back within 48 hours predicted as possibly not coming back at all. These customers should get an offer after two days and we'd be missing them.

Class 1 predicted as Class 0 (left center on matrix): Customers who will make their next purchase in 2-5 days predicted as possibly not coming back at all.

Class 2 predicted as Class 1 (bottom center on matrix): These customers are getting offers, just three days late.

So, while we sacrificed a few accurate predictions in Class 2, our second model reduced our least acceptable error from 16 percent to 3 percent and our second-least acceptable error from 43 percent to 20 percent.

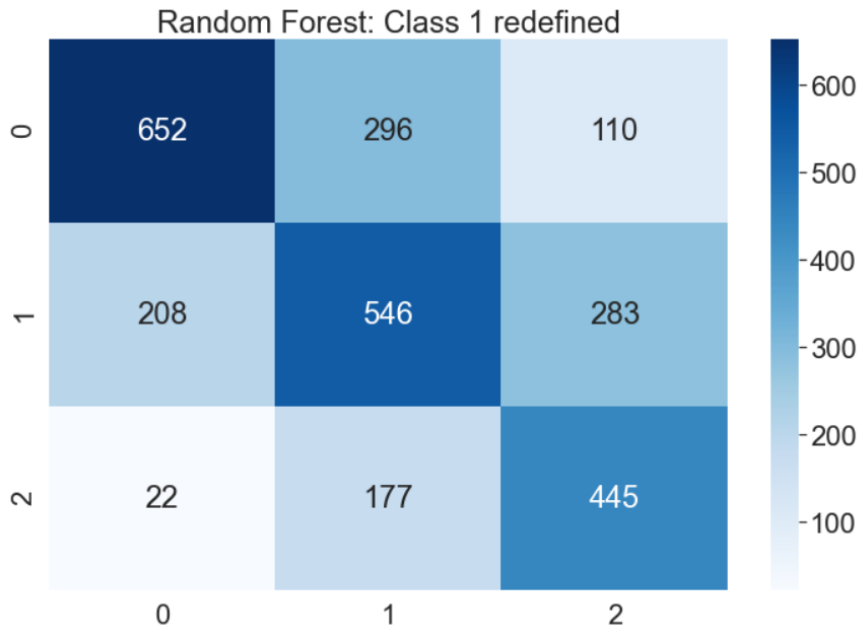
The first model is actually slightly more accurate overall (.606 to .5999), but it's significantly more accurate when it comes to Class 2 and Class 1, our most important classes.

They're people, not percentages

Let's now look at the same data, except instead of percentage points we'll use the raw numbers. Our test set, which was used as the final evaluation for each model, contains 2,739 customers.

We have 644 Class 2 customers in our test set, and we have our eye on 622 of them. Of those 622, 445 will get an offer after 2 days if they haven't purchased again, and the other 177 will get an offer after 5 days if they still haven't purchased again.

We have 1,037 Class 1 customers and 829 of them, if they haven't purchased again in 5 days, will get an offer.



Conclusion

The chart below clearly demonstrates our model's most powerful feature.

The more recently the customer has made a purchase, the more likely that customer is to come back sooner. The longer the customer stays away, our chances of losing that customer increase.

This underscores the importance of getting those offers out if a Class 2 customer hasn't returned in 48 hours, or if a Class 1 customer hasn't returned in 120 hours.