Original Case Study Narrative – Part 3
Rebecca Lewis
DSC 550
May 3, 2020

Case Study: Analyze survey data to predict how likely a client is to return to the SPCA community clinic

Data File

New community clinic clients at the Louisiana SPCA are sent a survey to determine if they were satisfied with their service and give them an opportunity to provide feedback. The survey contains the following data:

1. Multiselect response for how they heard about the clinic.
2. Single selection option for the type of service they had. In the event "other" is selected, they can enter more information
3. A rating from "Extremely Likely" to "Not Likely At All" on whether they would return to the clinic.
4. A free from response where additional information can be provided

Hypothesis

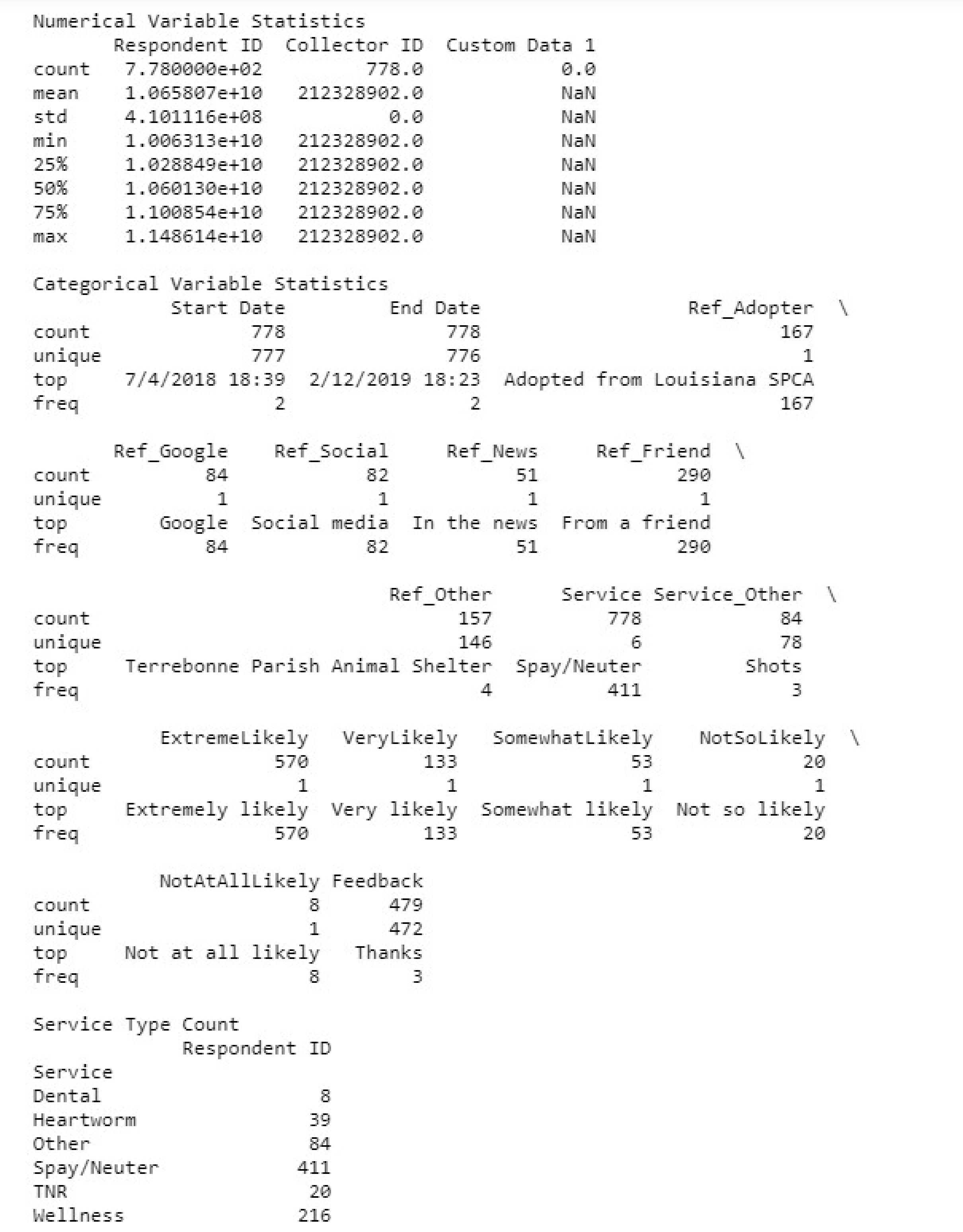
My hypothesis is that clients who have adopted from the Louisiana SPCA already are more likely to have a good experience and are more likely to return. For those who are not likely to return, is there any information provided in the additional feedback response that can provide information on areas of improvement.

Graph Analysis Instructions

Before reading in the data and working in Python, I removed any identifying data from the file for the client including name and contact information. A unique numerical identifier remains.

1. Read in data from the "SPCA_Clinic_Survey_Redacted.csv" file.
2. Clean up headers and other data for clearer summarization and graph analysis. Steps taken include:
    a. Headers are contained in both the first and second rows. The first row contains the main header/question. The second row contains the response options for multiselect and drop-down responses. Keep appropriate headers from row one and replace headers from row two if needed. Then remove the second row from the data.
3. Review the summary data.

Original Case Study Narrative – Part 3
Rebecca Lewis
DSC 550
May 3, 2020

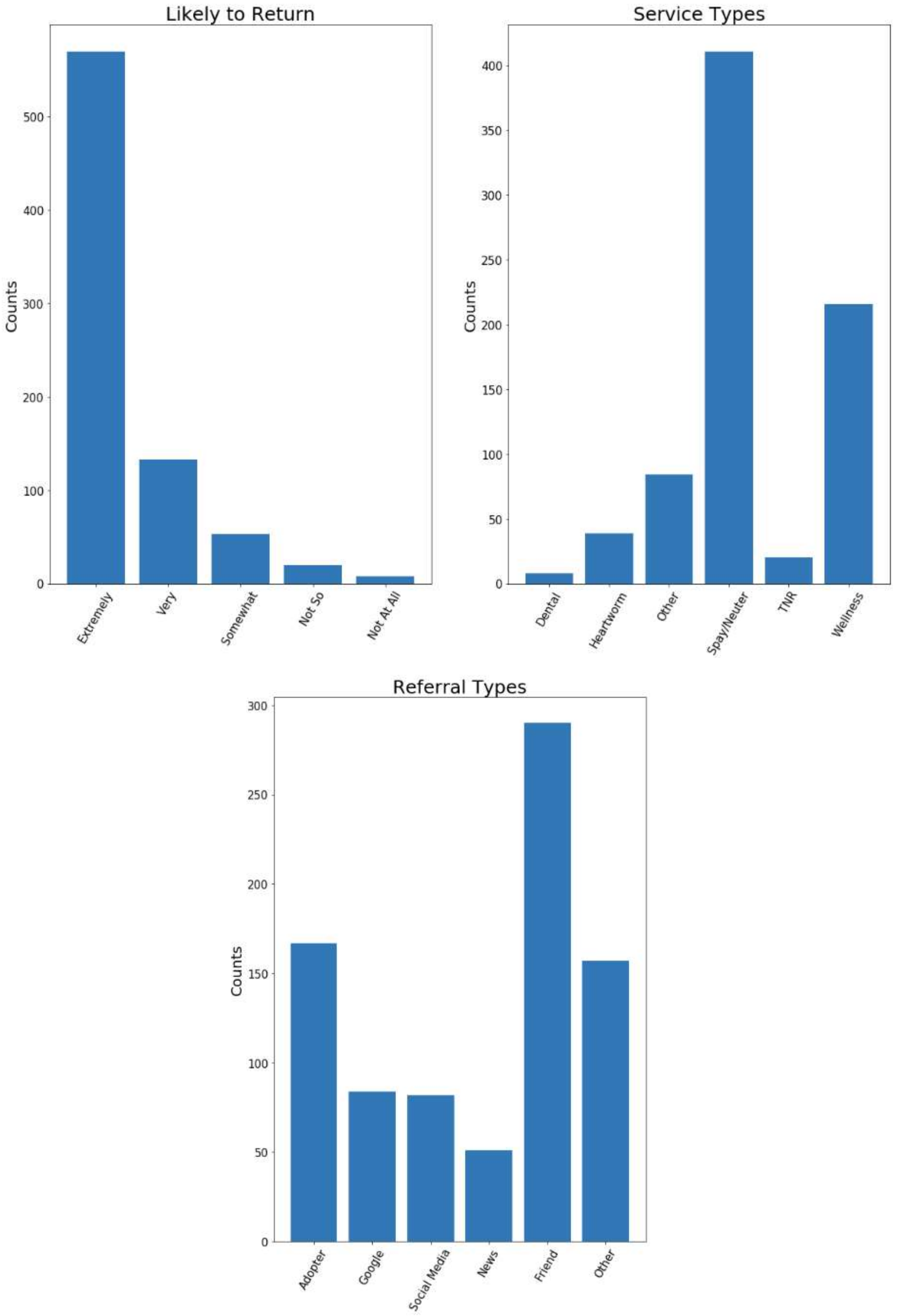


```
Numerical Variable Statistics
       Respondent ID  Collector ID  Custom Data 1
count   7.780000e+02         778.0            0.0
mean    1.065807e+10   212328902.0            NaN
std     4.101116e+08           0.0            NaN
min     1.006313e+10   212328902.0            NaN
25%     1.028849e+10   212328902.0            NaN
50%     1.060130e+10   212328902.0            NaN
75%     1.100854e+10   212328902.0            NaN
max     1.148614e+10   212328902.0            NaN

Categorical Variable Statistics
             Start Date         End Date                  Ref_Adopter  \
count               778              778                          167
unique              777              776                            1
top     7/4/2018 18:39  2/12/2019 18:23  Adopted from Louisiana SPCA
freq                  2                2                          167

       Ref_Google    Ref_Social     Ref_News     Ref_Friend  \
count          84            82           51            290
unique          1             1            1              1
top        Google  Social media  In the news  From a friend
freq           84            82           51            290

                              Ref_Other      Service Service_Other  \
count                               157          778            84
unique                              146            6            78
top     Terrebonne Parish Animal Shelter  Spay/Neuter         Shots
freq                                  4          411             3

           ExtremeLikely   VeryLikely  SomewhatLikely    NotSoLikely  \
count                570          133              53             20
unique                 1            1               1              1
top     Extremely likely  Very likely  Somewhat likely  Not so likely
freq                 570          133              53             20

           NotAtAllLikely Feedback
count                   8      479
unique                  1      472
top     Not at all likely   Thanks
freq                    8        3

Service Type Count
             Respondent ID
Service
Dental                   8
Heartworm               39
Other                   84
Spay/Neuter            411
TNR                     20
Wellness               216
```

Original Case Study Narrative – Part 3
Rebecca Lewis
DSC 550
May 3, 2020

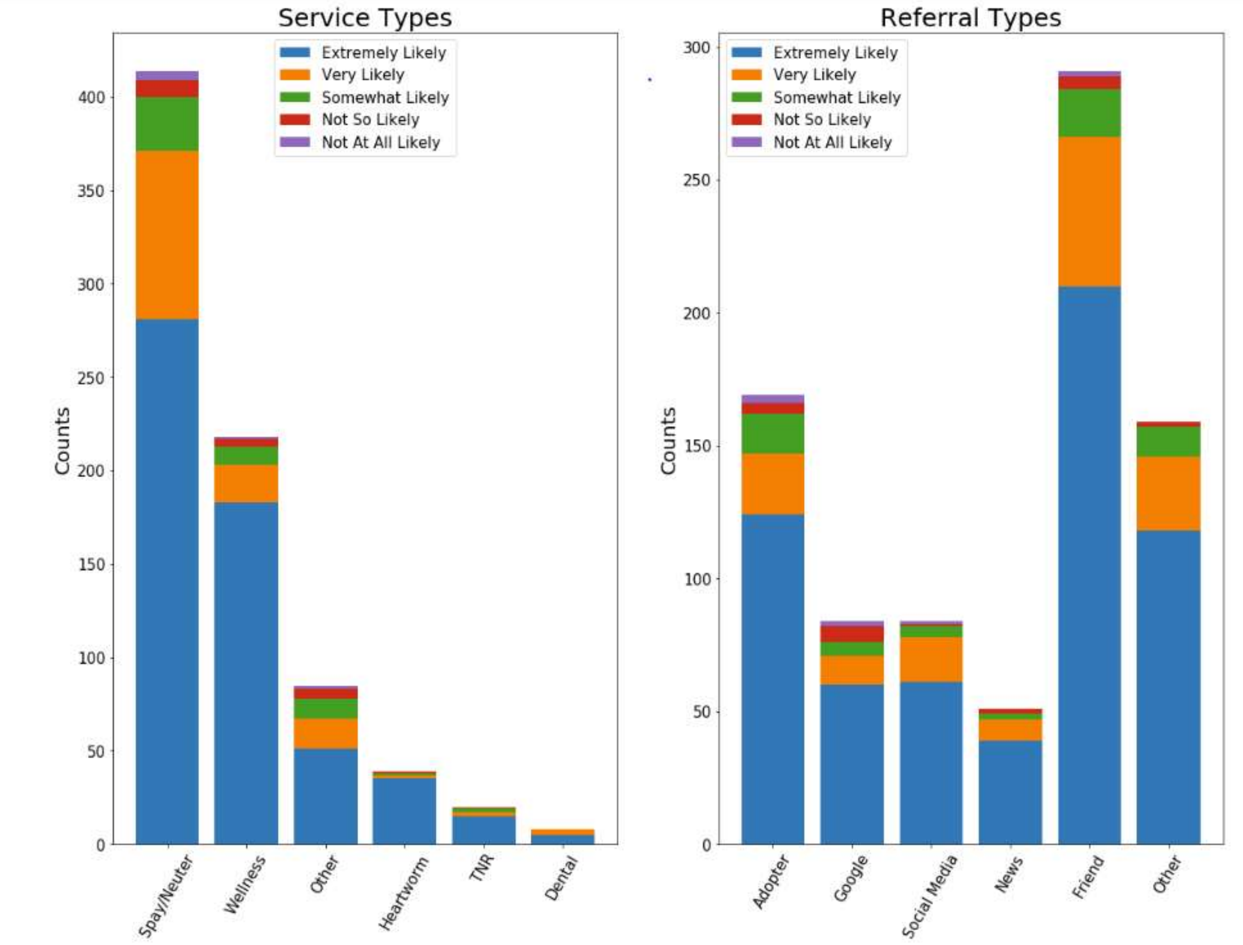
4. Prepare the data for Graph Analysis:
    a. For the categorical variables contained in separate columns, I applied one hot encoding replacing missing values with 0s and non-null values with ones. I kept a copy of the free form notes.
5. Display the formatted data before beginning analysis.
6. Form questions about the data and make observations.
    a. The target for training the model will be the Retention Ratings transformed to a single column in numerical form. The features will be the referral type and the service type.
    b. Overall most clients state they are likely to return. We can review to see if any of the categorical variables influence that decision: by referral type and service
    c. There are 6 unique service types however when I see the top service for other, it states shots. Shots are considered wellness. So some text analysis and recategorization could be considered to determine if a better service category exists.
    d. How will those who answered Somewhat Likely need to be handled? Should they be excluded so we can get a clearer picture of those who would return versus those who would not?
    e. For people who stated they did not want to return, are any keywords in their feedback that could indicate any potential challenges that could be resolved to improve their client retention.Plot the data with Bar Charts to view the distribution of clients among the categories and note additional observations.
7. Plot the data with Bar Charts to view the distribution of clients among the categories and note additional observations.
    a. I did not expect to see that most people were referred by friends. I'm interested to see the percentage of those who were referred by friends who said they would return

Original Case Study Narrative – Part 3
Rebecca Lewis
DSC 550
May 3, 2020



Original Case Study Narrative – Part 3
Rebecca Lewis
DSC 550
May 3, 2020

8. Plot the data with Stacked Bar Charts to observe the distribution of responses among the feature variables and note additional observations.
    a. It appears that most people who will not return are located in the Spay/Neuter service type group; however, that group also had the most visits period so we would need to look at the percentage of those who returned in each group to compare.
    b. The referral type that had the most people who stated they would not return was Google. This had one of the fewest people which could be meaningful. Perhaps the google search criteria needs to be tweaked as people are not getting what they were expecting.

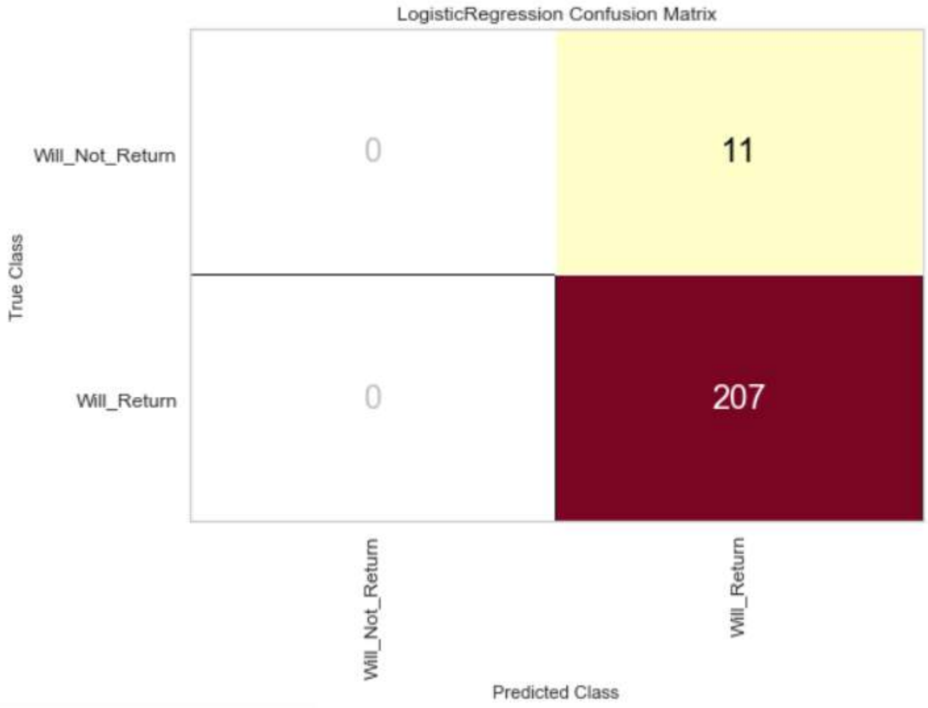


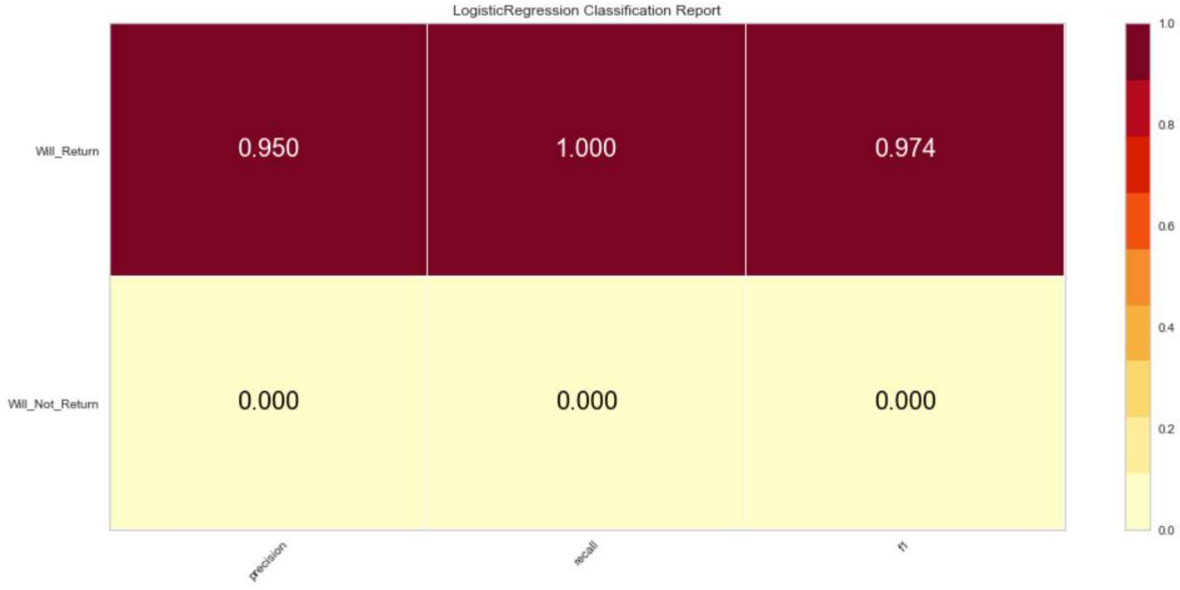
    a. Consolidate ratings into fewer categories that represent whether or not the user will return. Extremely Likely and Very Likely were mapped to yes. Not So Likely and Not Likely at all were mapped to no. Somewhat likely was mapped to Neutral until further information can be gathered to determine if these would be considered more yes or no.

9. Drop columns not needed for the analysis
10. Create a target column that groups together the retention ratings that make sense.
    a. Extremely likely to return, and very likely to return will be considered Yes
    b. Somewhat likely to return will be considered Neutral. A decision needs to be made on whether to group these responses into the yes or no category. I'm currently taking the approach to just remove them from the analysis until I can hear back from the SPCA Communications Director.
    c. Not So Likely and not at all likely will be No.
11. Convert any remaining categories to numerical representation.
12. Evaluate the data in the feedback provided from the Other column to determine whether any meaningful service categories could be derived from text analysis. Most seem to be related to a service type but have an exception such as the service ended up being canceled or they had more than one service type per visit. Ultimately, I chose to keep this variable as is and not make any assumptions about the respondent's answer to maintain the integrity of the data.
13. Split the data into testing and training feature and target datasets. The Louisiana SPCA Communications Director wants to leave out the neutral responses so those were eliminated from the target dataset.
14. Choose a learning algorithm, fit a model and view the evaluation metrics.:
    a. Logistic Regression was chosen initially because the features and target were all binary values. The score for the model was 94% but the model was classifying all of the observations as will return. Because there were not many will not return responses, imbalanced classes were most likely causing the issue. I modified the model to set class_weight = "balanced". Overall, the score decreased but the amount of false positives were reduced which was an improvement; however, this change also resulted in some false negatives.

Original Case Study Narrative – Part 3
Rebecca Lewis
DSC 550
May 3, 2020

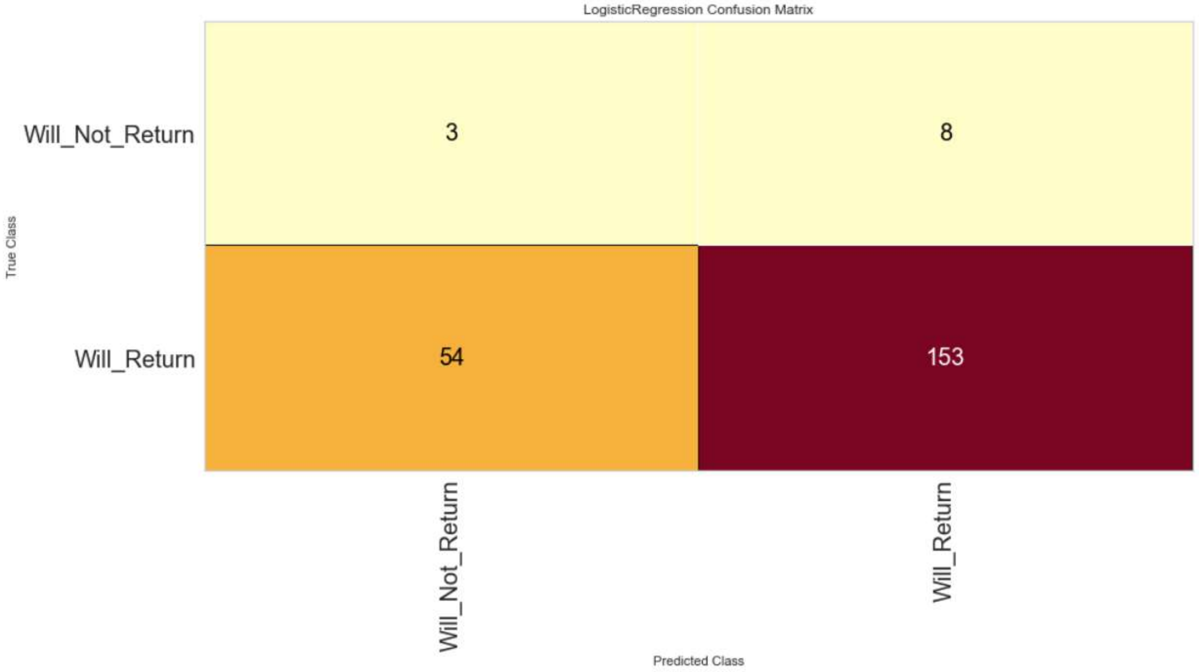
Logistic Regression Performance Before Weighting Classes

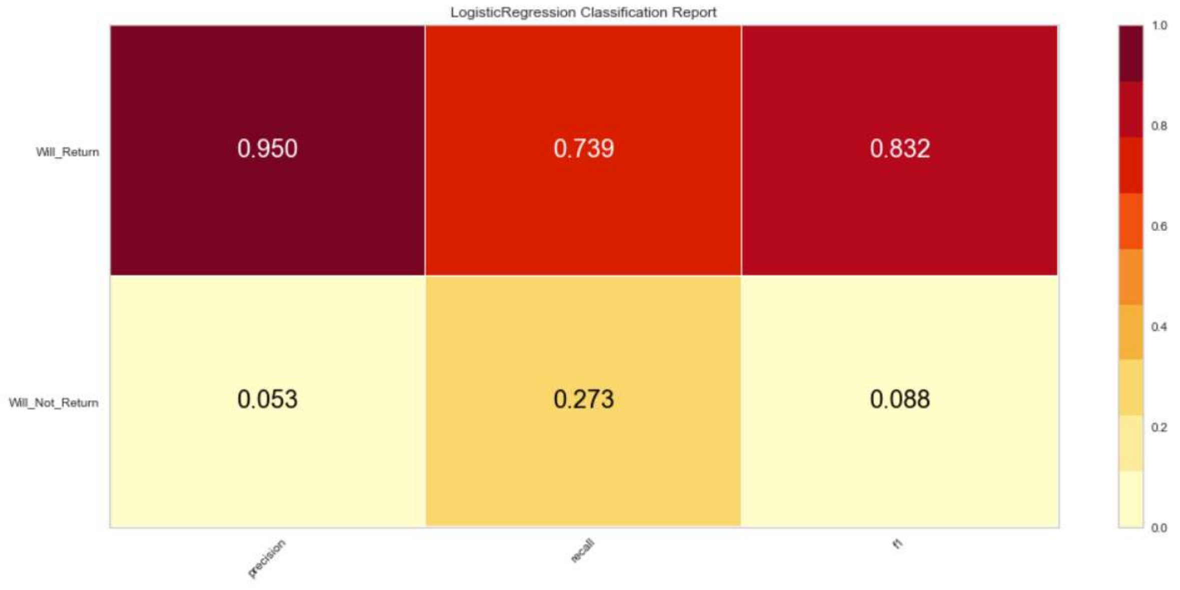




Original Case Study Narrative – Part 3
Rebecca Lewis
DSC 550
May 3, 2020

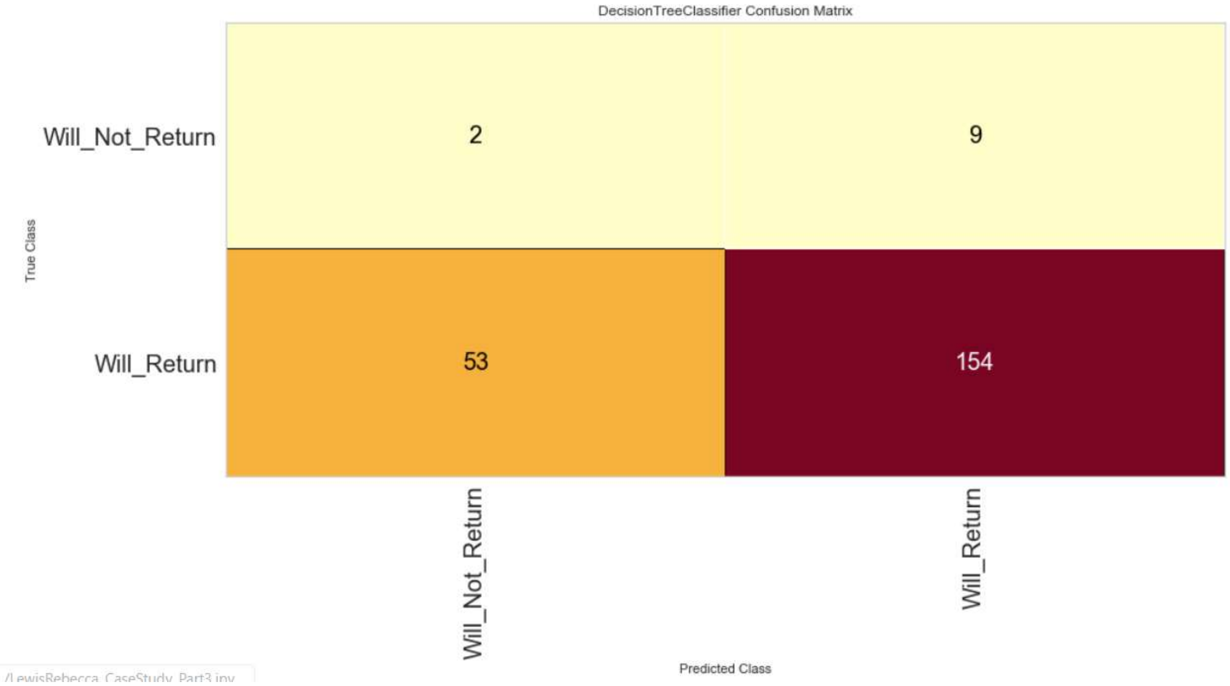
Logistic Regression Performance After Weighting Classes

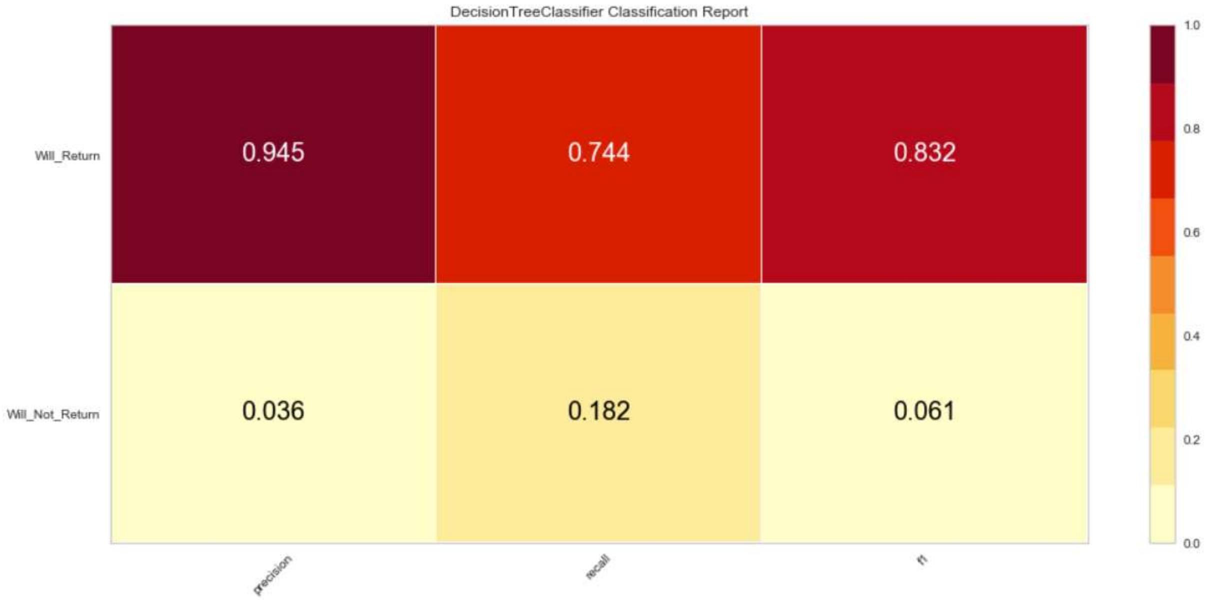




b. Decision Tree Classifier was attempted next with class_weight set equal to balanced. It produced similar results to the weighted Logisitic Regression Model. The weighted logistic regression model was slightly better at predicting people who were not likely to return.

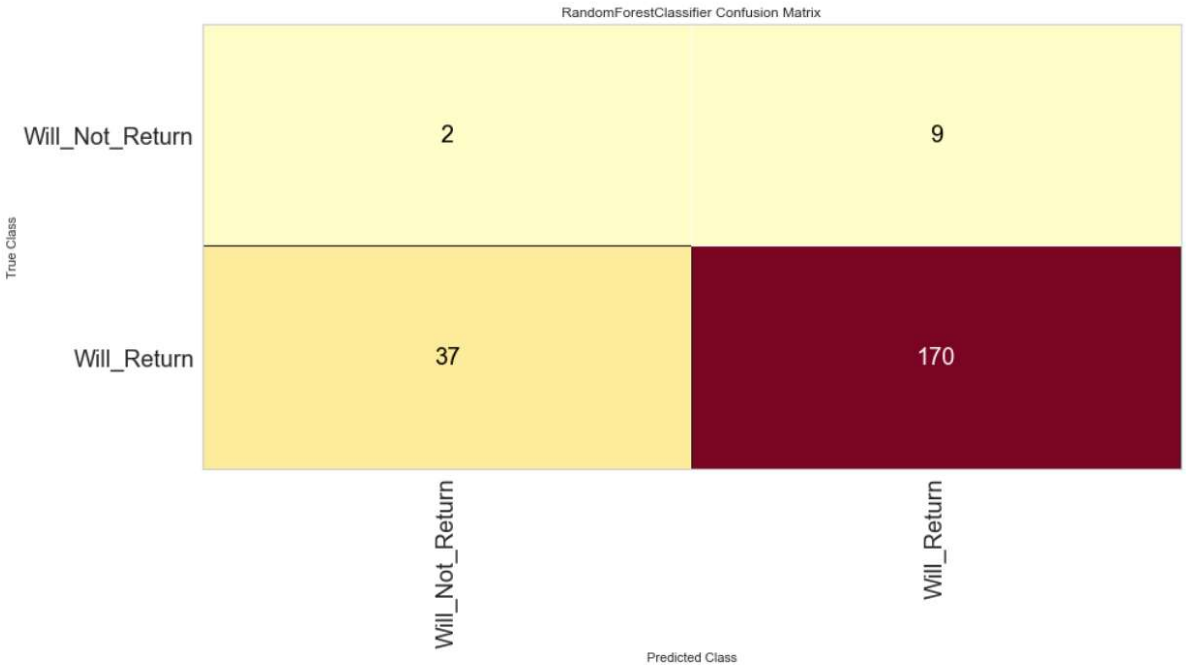
Decision Tree Classifier Performance

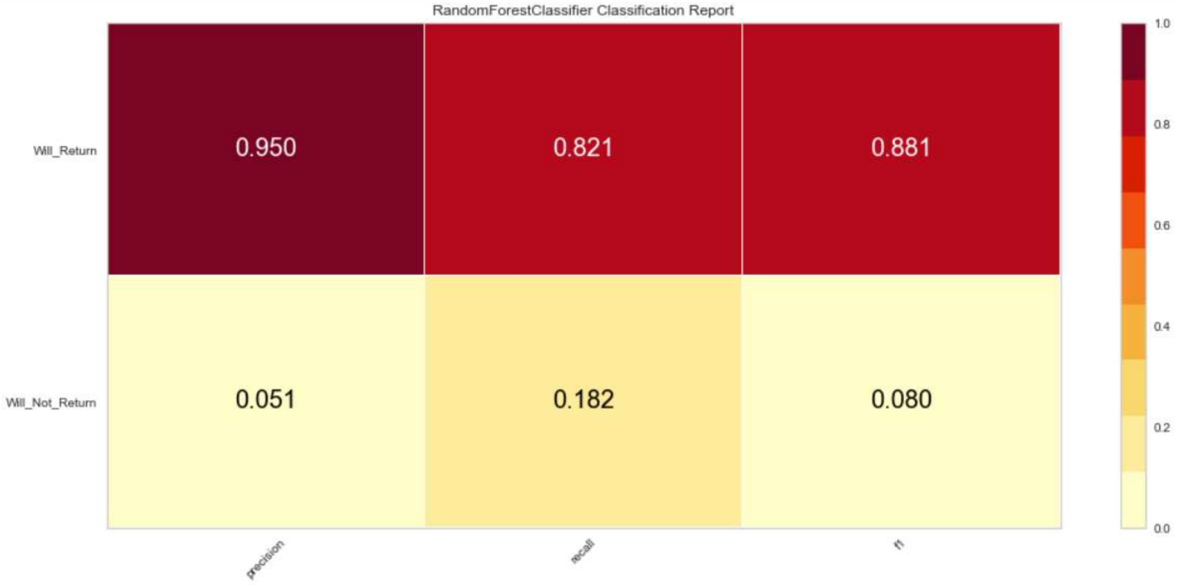




c. A weighted Random Forest Classifier was applied to the data and produced the best performing model as of yet. The precision, recall and f1 score increased over the decision tree classifier for those who stated they will return and the precision and f1 score for those who said they would not return also increased. The logistic regression model with balanced classes is still the best at predicting who will not return.

Random Forest Classifier Performance

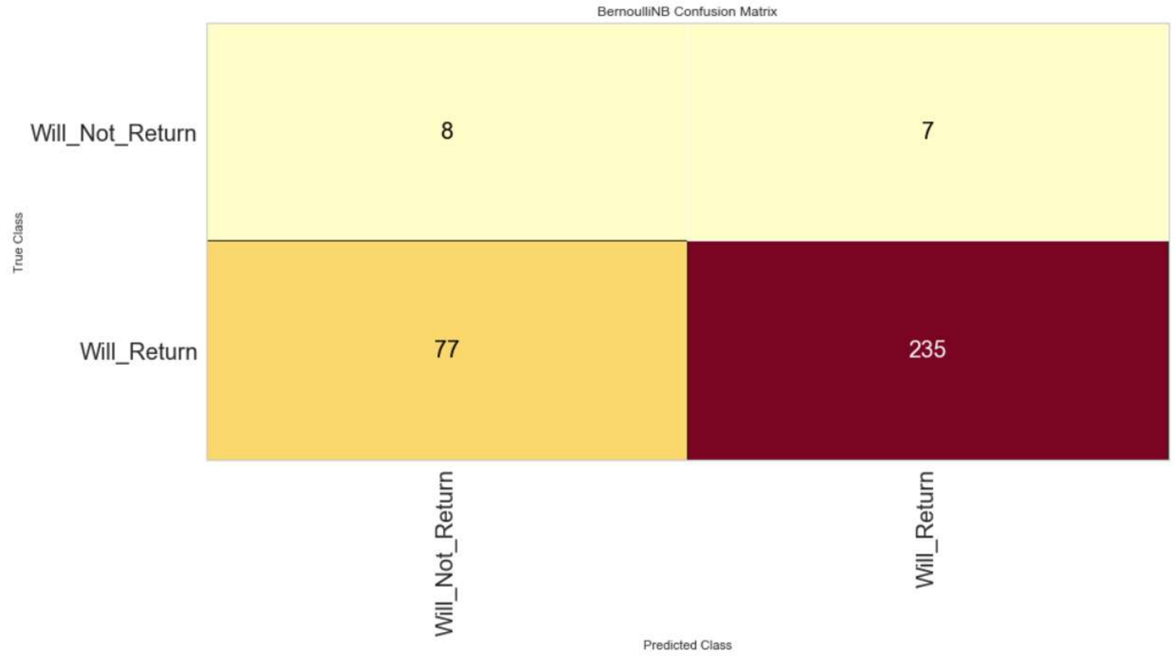


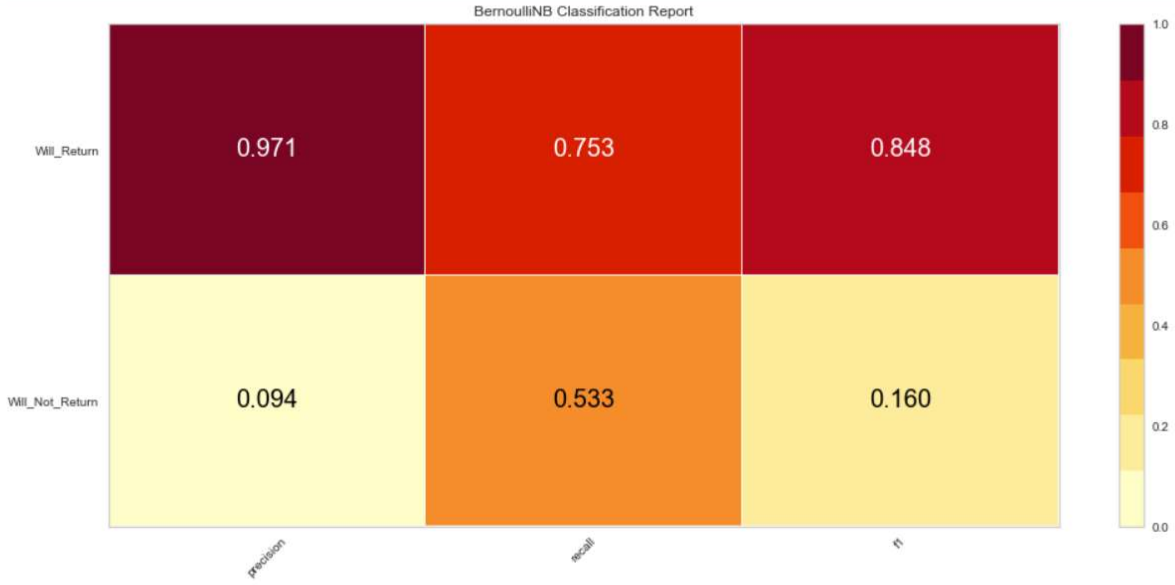


Original Case Study Narrative – Part 3
Rebecca Lewis
DSC 550
May 3, 2020

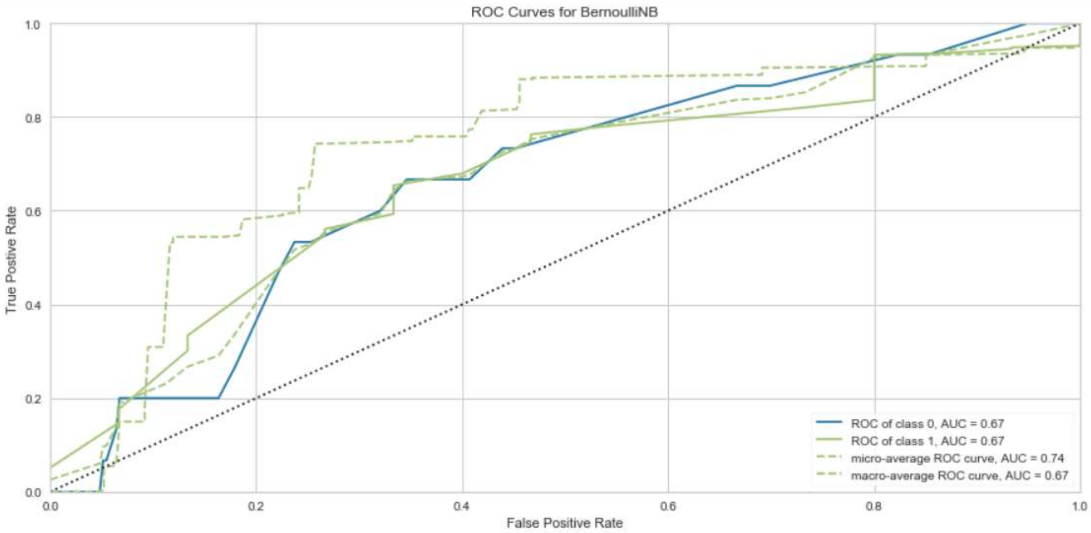
d. Naïve Bayes classification at first look seemed to produce a worse performing model as the confusion matrix was lower but when looking at the classification report, it seems to have the best results for precision, recall and f1 score for both classes.

Naïve Bayes Classifier Results

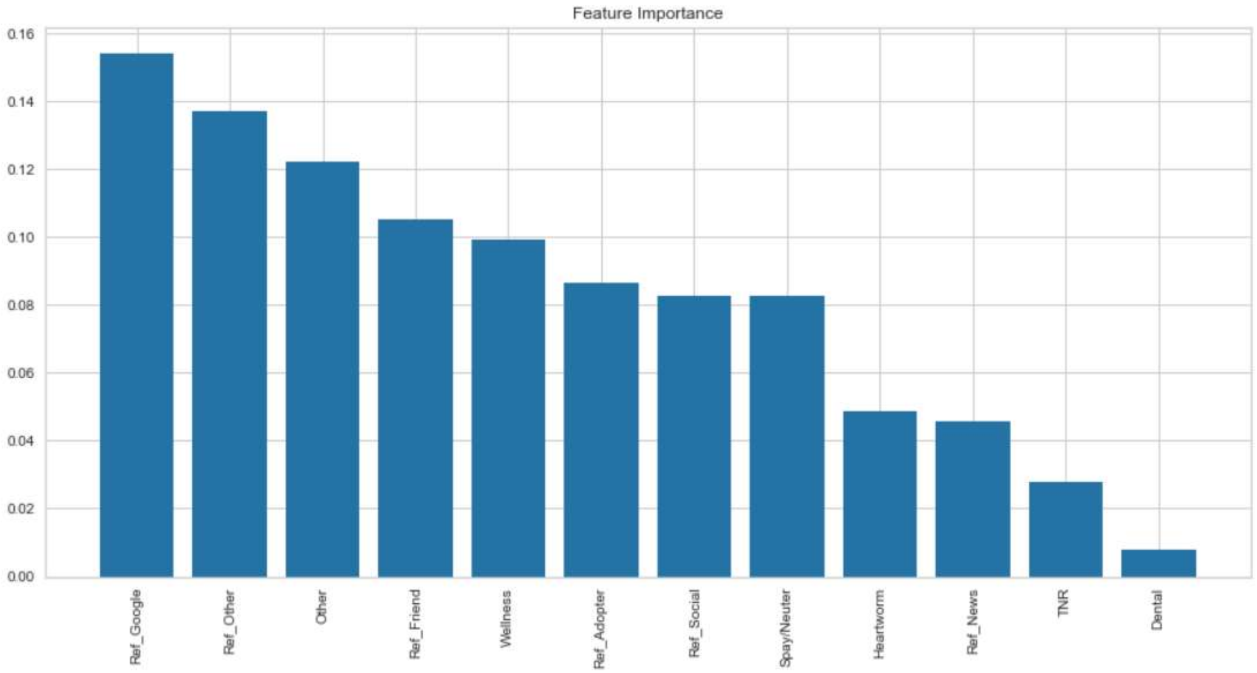




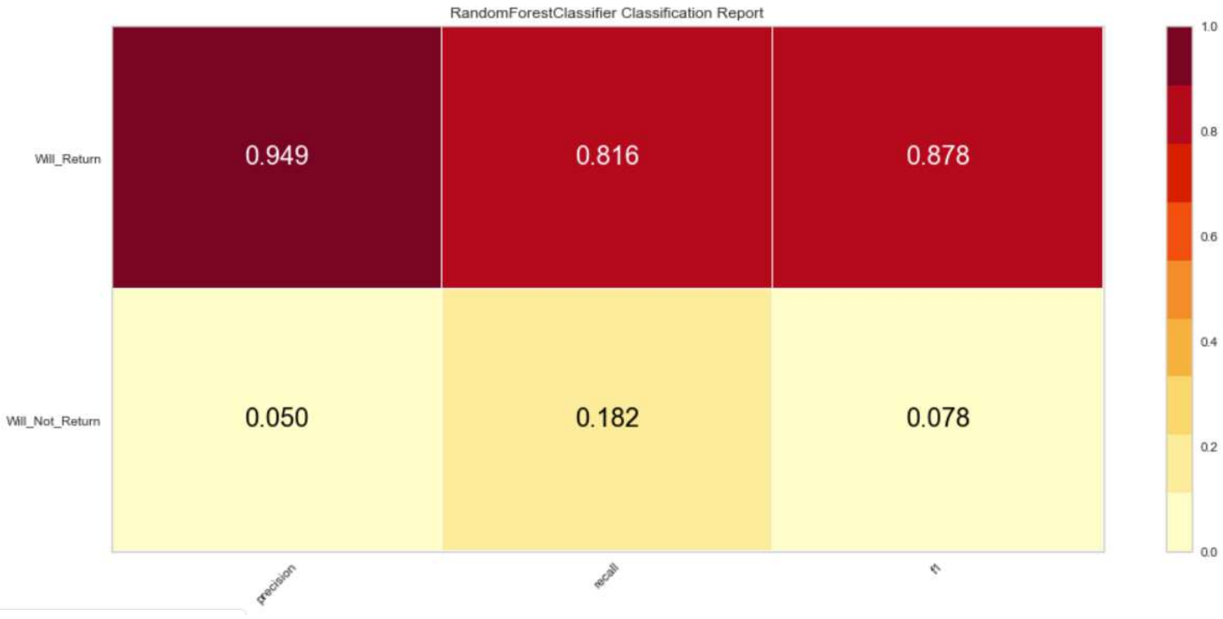
Original Case Study Narrative – Part 3
Rebecca Lewis
DSC 550
May 3, 2020



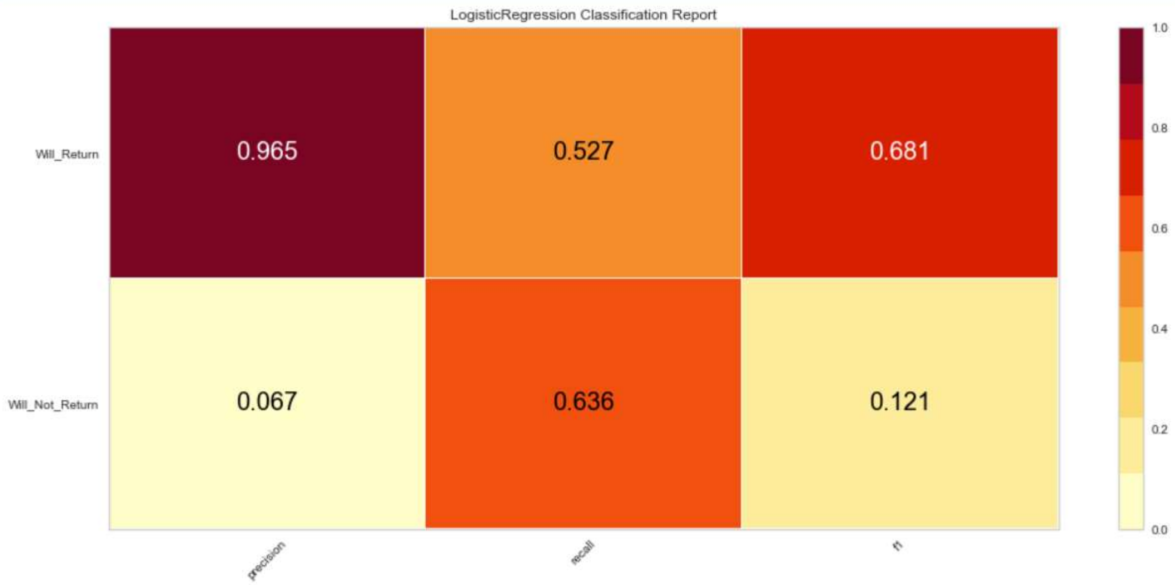
15. Evaluate the features used in the model to see if further feature selection could improve results.
    a. View feature importance and rerun the algorithm based on a threshold that only includes the important features. There was a significant gap in importance around .08. I refitted the model and validated the test data using only the features with an importance above that threshold. Based on the metrics, there was no significant improvement in the model.



Original Case Study Narrative – Part 3
Rebecca Lewis
DSC 550
May 3, 2020

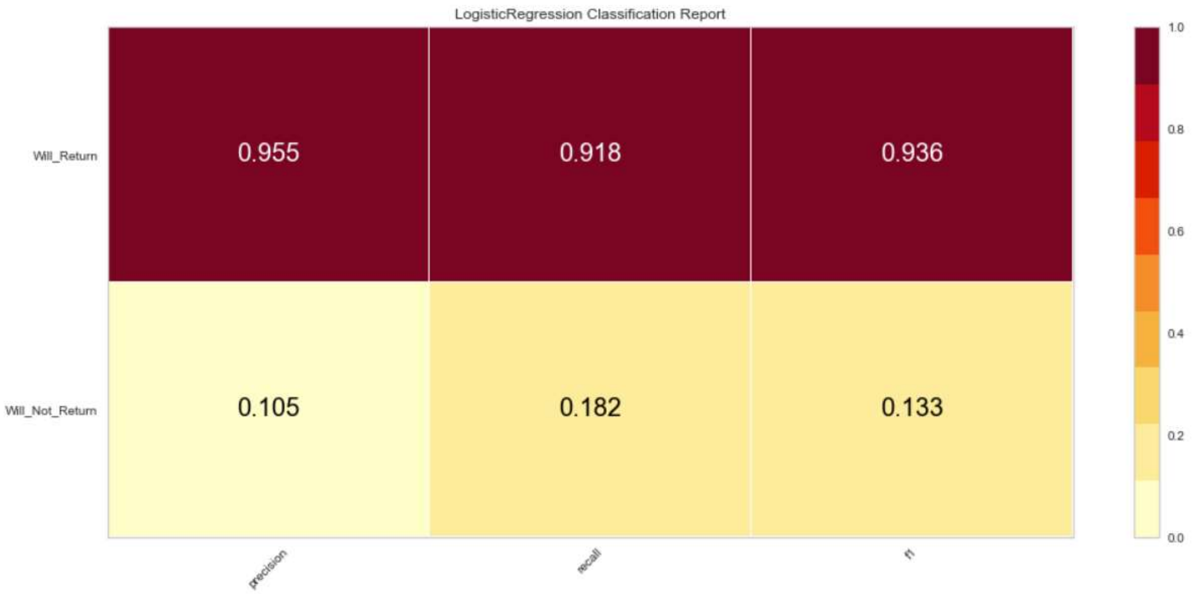


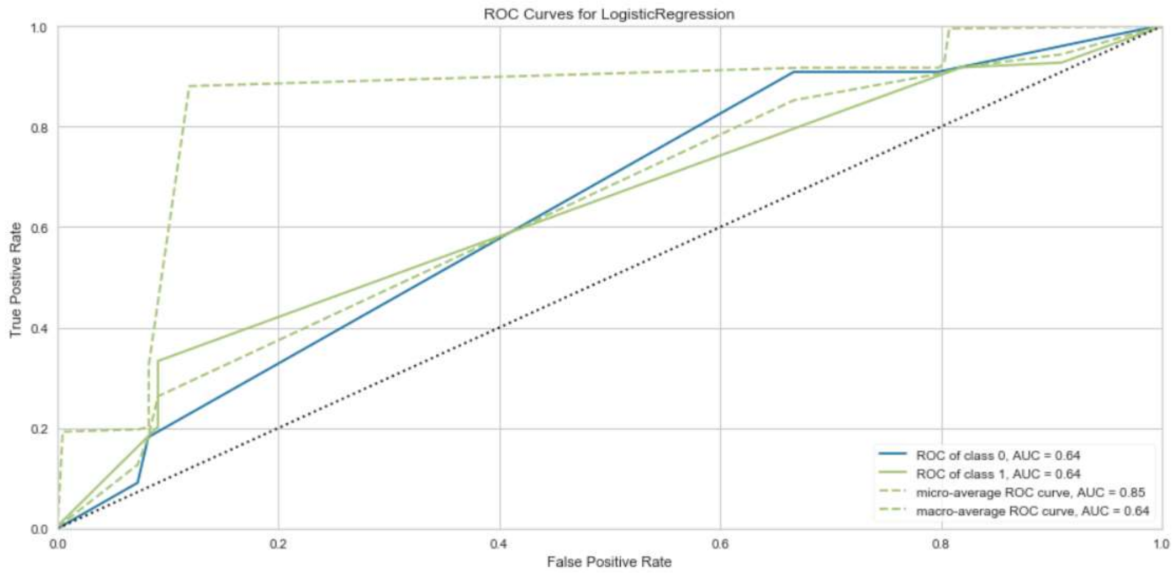
b. Variance Thresholding returned Ref_Friend, Spay/Neuter, and Wellness as the variables with the highest variance. I created a new logistic regression model with just these three features and received better results for people who stated they would not return. The performance for people who stated they would return decrease but the f1 score is still closer to 1 than 0:



c. Recurring Feature Elimination with Cross Validation was used with Logistic Regression, Random Forest Classifier and Naïve Bayes.

   i. For logistic regression, the final features were Ref_Social, Ref_Other and Other Service Types. Retraining the model with only these features results in improved performance of those who will not return without negatively impacting the performance of predicting who will return as much as Variance thresholding did.

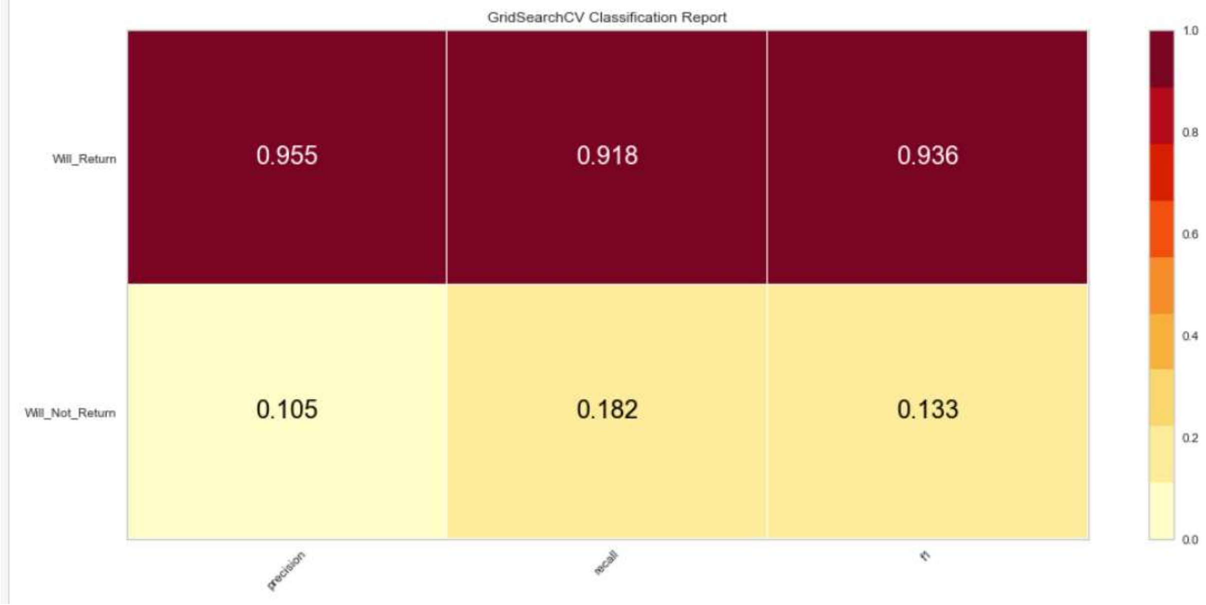




   ii. For Random Forest Classifier, the best features were the same features that were identified through the feature importance process in 15a.

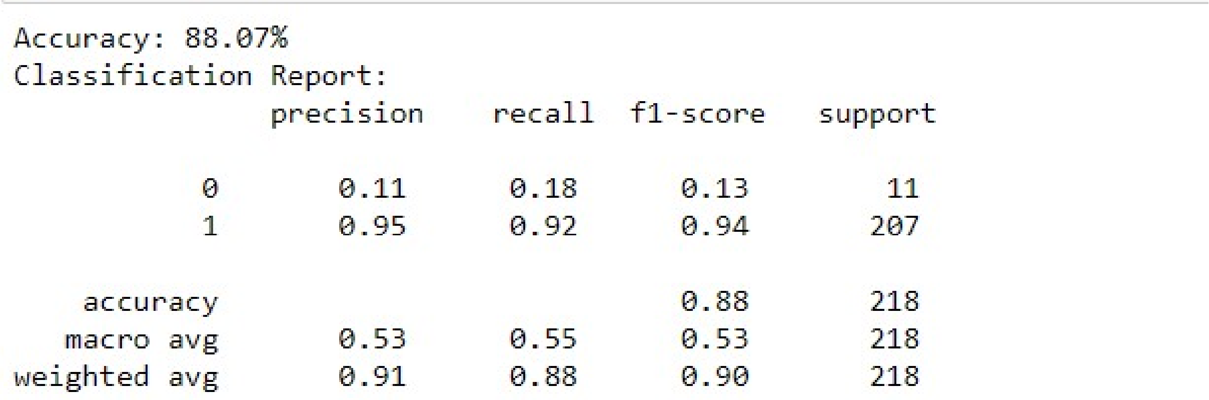
    iii. Using recursive feature elimination with Naïve Bayes returned the Dental column. Because none of my other analysis resulted in Dental as an important feature, I'm assuming this does not work well for Naïve Bayes for this particular dataset.

16. Use GridSearchCV to find the best hyperparameter values. A penalty of l2 and C of 1 were determined the best. After retraining and retesting the model, the results did not improve.



17. Analyze the comments provided by people who stated they would not return to determine any common themes. Using Tfidf vectorizer for a max of 10 features revealed that 'wait' and 'time' were ranked at the top of the list. Knowing what I do about the community clinic, this leads me to believe that long wait times may have contributed to their desire to return to the clinic.
18. Overall, the best model for predicting the likelihood that someone would return to the community clinic is using balanced logistic regression using the Social Media Referrals, Other Referrals, and Other Service types. (identified in step 15cii) Additional observations based on these results:
    a. The Louisiana SPCA has an excellent social media presence which would make sense why it influences clinic clients.
    b. More analysis would need to be performed on the customers notes for other referrals to gain an understanding at how this influenced the results.
    c. Other service types could include emergencies where the vet and staff went above and beyond to help the animal or could be the result of an unsatisfactory outcome for the animal. More analysis could be performed on the notes for other service types to see how they influence the retention.
    d. The model is 88% accurate overall, most likely because the classes are highly imbalanced and the model is much better at predicting those who will return. This case study could be

repeated with an equal number of likely to return responses and not likely to return responses using random sampling once more responses are collected. Currently, there are only 28 “not likely to return” and the dataset would be too small to get meaningful results.



```
Accuracy: 88.07%
Classification Report:
              precision    recall  f1-score   support

           0       0.11      0.18      0.13        11
           1       0.95      0.92      0.94       207

    accuracy                           0.88       218
   macro avg       0.53      0.55      0.53       218
weighted avg       0.91      0.88      0.90       218
```