Clark Farnsworth
Stockton Smith
CS 5830
Project 7

Link to github: here
Link to slides: here
Link to airplane dataset: here
Link to shopping dataset: here

Introduction:

In this report, we explore the process of creating a logistic regression model capable of predicting whether or not airline passengers are satisfied with their flight experience. The stakeholders for this analysis are airports and airline companies, as our results will allow them to know which parts of the travel experience are most important to passengers, enabling them to focus their efforts in those areas to improve passenger satisfaction.

Additionally, we developed a logistic regression model aimed at discerning whether an online shopper completed a purchase on a shopping website. This endeavor is significant to online retailers (the stakeholders for this analysis), as their primary objective is to maximize conversions, thereby transforming visitors into customers. By delving into the user behavior, this model unveils crucial insights into the determining features that drive purchasing decisions. Such insights empower retailers to refine their strategies and optimize their online platforms, ultimately enhancing their profitability and competitiveness.

Dataset:

The airline passenger dataset contains several columns describing a passenger's flight experience (typically measured on a scale of 1-5 based on survey responses), as well as a binary target variable indicating whether or not the passenger was satisfied with the experience overall. In order to clean and prepare this dataset for analysis, we first found it necessary to remove all entries with answers of zero (meaning 'not applicable') in survey question columns. We also created a new "satisfaction_bool" column in the dataframe, which simply changed values of "satisfied" to "True" and "neutral or dissatisfied" to "False". These changes were made so that the matrix and logistic regression analyses could be performed on valid inputs.

The online shopper dataset had mostly quantitative features, as well as 2 categorical features that we didn't find relevant for the analysis. Many of the features were from Google analytics. Some of the main features included exit rate, bounce rate, page values, and traffic type. One of the challenges we had with this dataset during preprocessing was that there were many more false cases (10422) than true cases (1908). To address this issue, we took an undersampling of the false cases, allowing us to obtain more balanced and meaningful results.

Analysis Technique:

For the airline passenger dataset, we chose to use logistic regression as our analysis technique. This was a suitable approach because our data consisted entirely of quantitative values except for the target variable (passenger satisfaction), which was binary and categorical. We chose logistic regression over a support vector machine since our dataset was very large, consisting of over 100,000 entries, and runtime performance was extremely poor for an SVM on such a large dataset.

For the online shopping dataset, we also used logistic regression as our analysis technique, as it matched well with the dataset. This is because the dataset had mostly quantitative features, as well as a categorical target feature indicating whether or not a shopper had bought an item on the website. Again, runtime performance was significantly better when using a logistic regression model rather than a support vector machine.

Results (airline dataset):

For the airline passenger dataset, we first ran a simple test in which we used the answers to all 14 survey questions as predictors for passenger satisfaction. This model returned precision, recall, and f-scores of approximately 0.82 for the positive case (satisfied passengers) and scores of about 0.86 for the negative case (unsatisfied passengers).

However, in order to provide the most useful information to our stakeholders, we set out to build a logistic regression model that could determine exactly which survey questions were the most critical to receive high scores on. To do this, we generated matrices showing each answer combination for each of the 91 attribute-pairs, along with the percentage of passengers for each cell that reported being satisfied with their flight. We then used logistic regression to compute the decision boundaries for each matrix. Cells with blue squares represent the side of the decision boundary predicting passenger satisfaction:

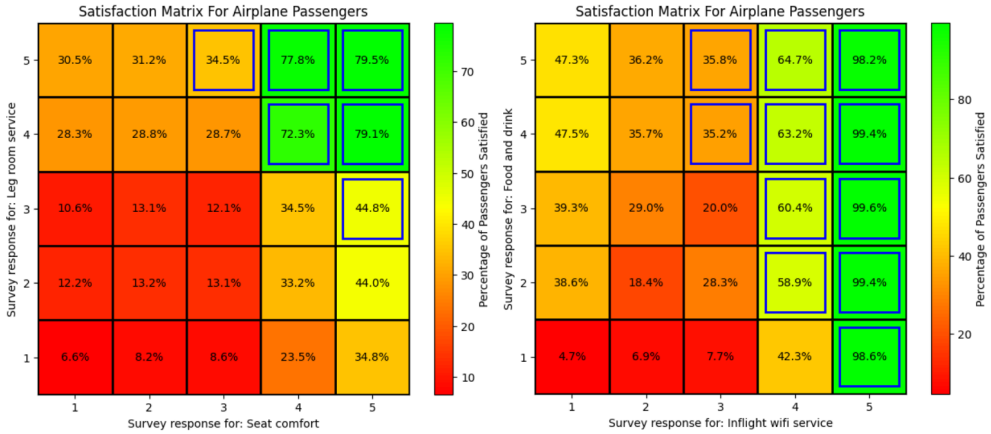

Fig. 1: Matrices using survey score pairs to predict overall flight satisfaction

Several interesting patterns can be noticed by analyzing these matrices; for example, it became clear that attributes such as “Food and drink” and “Gate location” are of little importance to passengers, as high satisfaction rates are typically reported (and predicted, as shown by the decision boundaries) even when these attributes are given the lowest possible score. Features such as “Baggage handling” and “Online boarding”, however, are critical in securing passenger satisfaction, as low scores on these questions almost always result in low satisfaction rates, even when other attributes receive perfect scores.

The decision boundaries computed by the logistic regression model are also significant in illustrating airline feature trade-off trends; for example, passengers are much more likely to compromise on “Ease of Online booking” if they receive better “On-board service” in return. These decision boundaries for the different attribute pairs were fairly successful in classifying passenger satisfaction correctly, returning the average scores shown below:

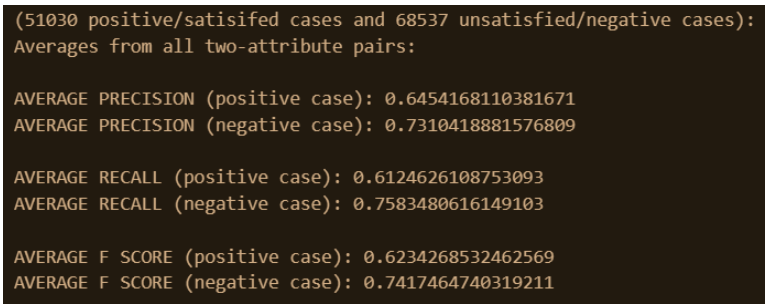

Fig. 2: Average precision, recall, and f-scores from the 91 different attribute pairs

The scores shown in Fig. 2 indicate that our two-attribute models performed nearly as well as our model which used all 14 attributes, with the benefit of illustrating more meaningful relationships than anything that could be learned from the all-attribute model. From this, we can conclude that our logistic regression models would be extremely useful for airports and airline companies in determining potential areas of improvement and trade-offs; this, in turn, would allow them to greatly enhance airline passenger satisfaction.

Results (shopping dataset):

Next, for the online shopping dataset, we began by running a logistic regression model with all 14 of the important features. For predicting the false case (that is, the case in which a visitor of the website did not end up making a purchase), the model computed a precision score of 0.777, a recall score of 0.886, and an f-score of 0.828. For the true case (the case in which a visitor *did* end up making a purchase), the model computed a precision score of 0.867, a recall score of 0.746, and an f-score of 0.802. As mentioned above, these scores were the result of undersampling the majority class (i.e. throwing out negative cases to ensure that our model would analyze an equal amount of positive and negative cases).

However, these results have the same issue as the original all-attribute results from the airline passenger analysis, which is that using all 14 of the features does not provide anything meaningful to the stakeholders of the websites. To improve the usefulness of the regression, we then paired up each of the different features and used logistic regression to predict the true and false cases. This allowed us to determine the pairs of attributes that had the highest success rates in predicting whether purchases were made by users. We also computed decision boundaries for every pair of attributes. Examples of these decision boundaries for different attribute pairs are shown in the figure below:

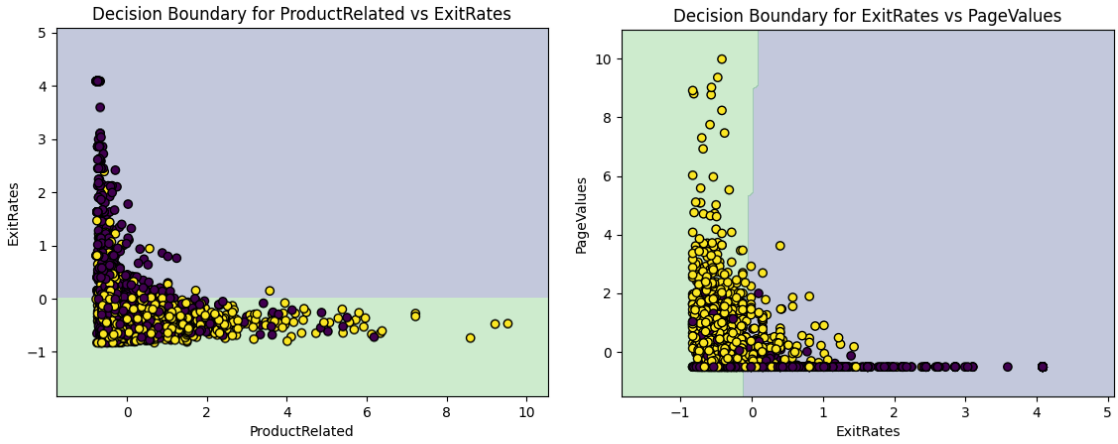

Fig. 3: Decision boundary plots for different attribute pairs. Yellow dots represent users who bought an item, and purple dots represent those who did not

After running logistic regression on every pair of features, we found that “page value” (quality of user experience provided by a page) and “product related” (how similar the purchased product was to other products viewed by the user) was the best attribute pair for predicting purchases, as the success rates were significantly high in both the positive and negative cases: for the positive case, the precision score was 0.871, the recall score was 0.735, and the f-score was 0.797. For the negative case, the precision score was 0.771, the recall score was 0.891, and the f-score was 0.826. These scores indicate that our final two-attribute model was just as effective (and in some cases, better) than the original all-attribute model. As a result, there is evidence to suggest that online retailers can generate higher profits as they seek to optimize their user experience and suggested product features on their websites.