Kieren Singh Gill, ksg7699
Responsible Data Science
Professor George Wood
Homework #2

Problem 1 (10 points): Online Job Ads

(a) (6 points) Give three distinct reasons why gender disparities might arise in the operations of such a system.

Pre-existing biases in the historical employment data: The machine learning system is trained with underlying historical data. Historically, males made up a higher percentage of higher-earning positions in the workforce due to gender bias in hiring, and these biases would be reflected in the data. Hence, because the machine learning model is trained on data that reflects this pre-existing historical bias, it will recommend higher-paying jobs to males, which will lead to gender disparities in the workplace as women and non-binary individuals will not be recommended these opportunities as frequently. Likewise, women will be recommended lower-paying jobs more frequently, which will perpetuate this gender disparity. Furthermore, there is a lack of historical data on the employment of non-binary individuals, simply because people weren't categorized as non-binary historically. This will result in the machine learning models recommending fewer jobs in general to non-binary individuals, simply because the models would have inferred that non-binary individuals have been less employed historically.

Black box models and Emergent bias from user interaction data: The machine learning system collects interaction data from its users, which is also used to train the model. This could result in the model becoming a black-box model, as the model could end up becoming such complicated functions of the variables that the people that designed the model fail to understand how the model comes to make its predictions. This makes it even harder to explain the decisions that the model makes, which means that it will be harder to understand why a model is discriminating against protected groups. In addition to this, if more males click on a particular job ad, the machine learning algorithm will learn from this pattern of user interaction and will recommend more jobs of this type to males. This is a form of emergent bias and will lead to a gender disparity within that industry.

Emergent biases from employer interaction data and financial optimization: Employers could have hiring biases, choosing to hire more males instead of females due to unfair reasons (such as the possibility for females to take maternity leave). Because the machine learning model is based on employer hiring data as well, the model could learn from this and become biased against hiring women. In addition to this, as discussed in class, Facebook's targeted ad system followed a distribution that was racially stereotypical. In this case, the online job ad system could follow a gender-stereotypical distribution, and the model could continue to recommend more jobs to males if they continue to click on these ads and get hired, as this means that the ad hosting service will receive more revenue.

(b) (4 points) Suppose that the job search service decides to increase the number of times it presents job openings in STEM to women. To do so, the service observes that STEM job experience (in years) is positively associated with the likelihood that a user clicks on an advertised STEM job opening: the more years of experience, the more likely a user is to click. Consider the following intervention:

> *Pre-process the training dataset, replacing the value of the "job experience" feature for women with the best (highest) possible value for the feature in the dataset.*

(i) Under what conditions will this intervention increase the number of times job openings in STEM are shown to women?

This intervention will increase the number of job openings in STEM for women if the "job experience" feature is considered of high importance in the machine learning model. As mentioned above, there is a positive association between job experience and user clicks on STEM job openings - the more years of experience, the more likely the person is to click. If the machine learning system is optimized for a high number of clicks on the job opening, since women are assigned the highest possible value of job experience in the dataset, the system will recommend STEM jobs to them. This will result in an increase in the number of STEM jobs shown to women.

(ii) Under what conditions will this intervention fail to increase the number of times job openings in STEM are shown to women?

This intervention will fail to increase the number of job openings in STEM for women if the "job experience" feature is considered of low importance in the machine learning model. If a different feature like "age" was given high priority, changing the value of job experience for women will have no impact on the number of job openings in STEM shown to women. Another scenario is that even if the "job experience" feature is considered of high importance for user clicks on ads, if the system optimizes the final hiring outcome instead of user click-through rate, this might not be as effective as there might be other factors that have a stronger positive association with someone being hired aside from job experience.

Problem 2 (20 points): AI Ethics: Global Perspectives

The Intersection of AI and Consumer Protection

In this lecture, Professor Jeannie Paterson discusses predictive analytics in targeted advertising and how it impacts consumer privacy in an adverse manner. There are many stakeholders involved in algorithmic advertising. Corporations have the opportunity to reach their ideal audience demographic through targeted advertising, which will help increase profits. Ad providers earn money because corporations pay them to run ads on websites. Websites that host these ads are also stakeholders, since the more ads on the site that people interact with, the more commission these websites receive from the ad provider. This incentivizes websites to allow cookies and web beacons on their platforms, and to curate content on their sites based on their user data to increase user engagement on their sites, as more user engagement results in more ad interaction. Policymakers are also stakeholders, as they are acting for the benefit of users to implement data privacy laws. Lastly, users face the tradeoff between the benefit of having relevant ads that are curated to their wants and needs and the downside of sacrificing their data privacy.

Websites use cookies and web beacons to collect and build user profiles. These profiles are used by websites and advertisers to provide targeted advertising to users. The issue with this is that most users are unaware of how advertising algorithms work, and this lack of transparency results in an information asymmetry that gives corporations and ad providers an unfair advantage. Professor Paterson spoke about how users are subject to differential pricing. This is when corporations show different prices to different users based on their user profile, which could include their demographic information and their spending habits, and this allows corporations to optimize pricing for a user based on their calculated propensity to spend. While this benefits corporations, this undermines a user's autonomy because in most cases they are unaware that they are being subject to this. Predictive analytics also undermines a user's autonomy because they will be shown a smaller range of products, undermining their ability for a wider choice of goods and services. Poorer people are also discriminated against, as corporations can display higher prices towards them for the purpose of discouraging them from purchasing their products. Even though more websites are reducing cookies as a means of protecting user data, this is insufficient, as web beacons are still implemented and are actively collecting user data.

There is currently legislation in the EU and in California that serves to protect user data by limiting what user data websites are allowed to track. However, these bills take a long time to be passed in court due to extensive lobbying by major software companies, and many other countries still lack these protections. The only incentive for corporations to be transparent would be to build trust with their consumers, as consumers are more likely to buy from firms that respect their privacy and follow ethical practices. An example of this transparency would be that many websites now have a cookie notification for first-time visitors.

Problem 3 (35 points): Generating Explanations with SHAP

(5 points) Use the provided Colab template notebook to import the 20 newsgroups dataset from sklearn.datasets, importing the same two-class subset as was used in the LIME paper: Atheism and Christianity. Use the provided code to fetch the data and split it into training and test sets. Then, fit a TF-IDF vectorizer to the data, and train an SGDClassifier classifier

Part (A)

```
[132] # Mark the categories of interest
      categories = ['alt.atheism', 'soc.religion.christian']

      # Fetch the data
      newsgroups_train = fetch_20newsgroups(subset='train', categories=categories)
      newsgroups_test = fetch_20newsgroups(subset='test', categories=categories)

      # Set outcome class names
      class_names = ['atheism','christian']
```

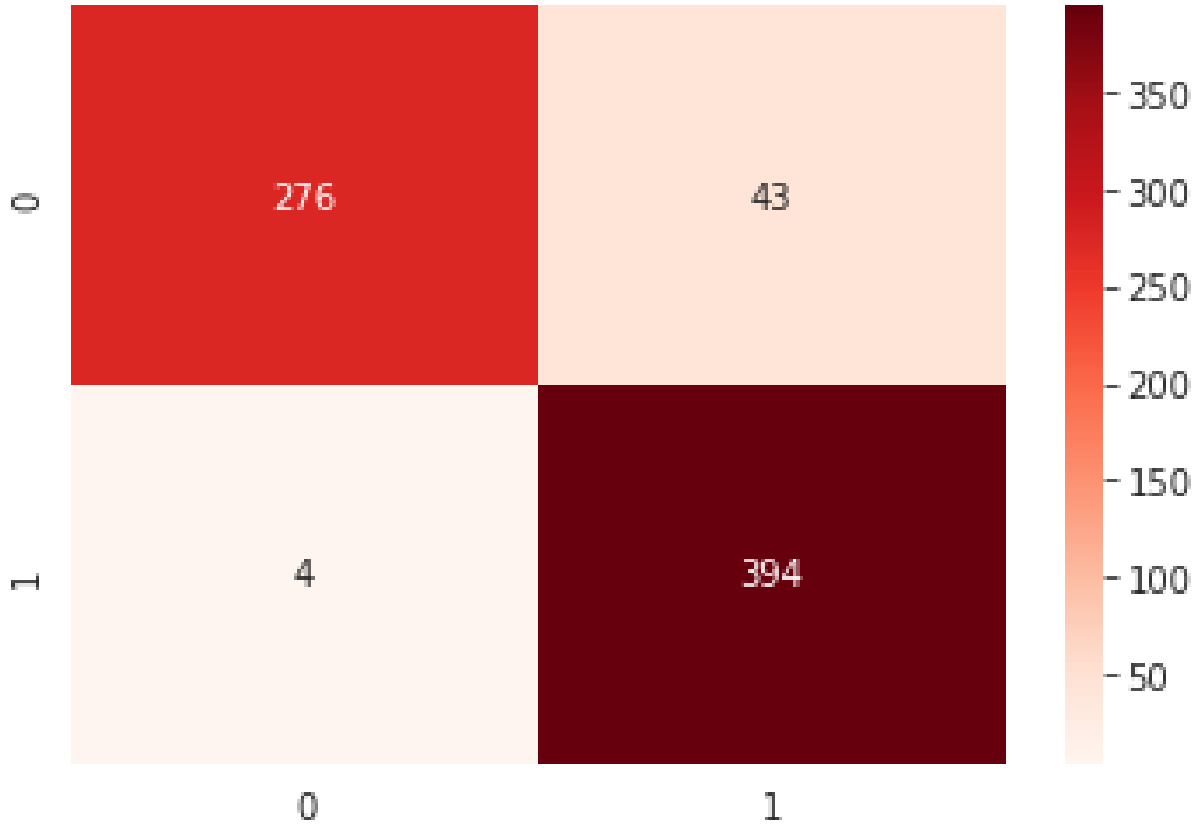
```
[133] # Initialize & fit tf-idf vectorizer (see notebook for lab 10)
      vectorizer = TfidfVectorizer(min_df=1)
      X_train = vectorizer.fit_transform(newsgroups_train.data)
      X_test = vectorizer.transform(newsgroups_test.data)
      Y_train = newsgroups_train.target
      Y_test = newsgroups_test.target
```

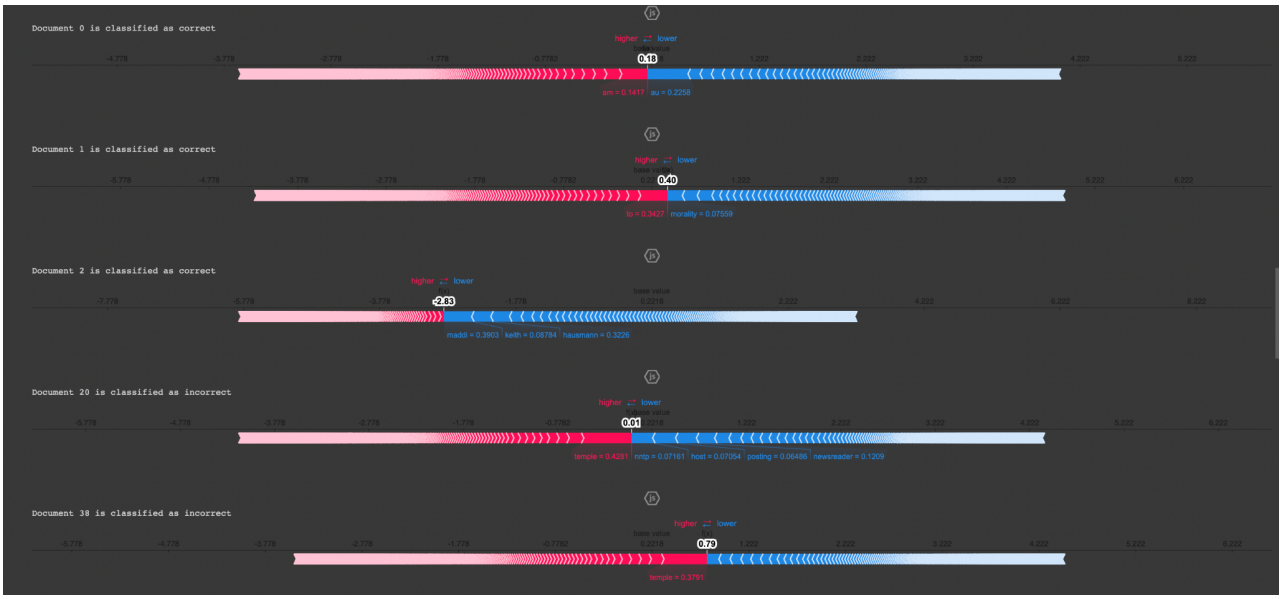
```
      # Train & fit the classifier
      model = SGDClassifier(loss="log")
      model.fit(X_train,Y_train)

      SGDClassifier(loss='log')
```

(b) (10 points) Generate a confusion matrix (hint: use sklearn.metrics.confusion_matrix) to evaluate the accuracy of the classifier. The confusion matrix should contain a count of correct Christian, correct Atheist, incorrect Christian, and incorrect Atheist predictions from your SGDClassifier. Use SHAP's explainer to generate visual explanations for any 5 documents in the test set. The documents you select should include some correctly classified and some misclassified documents

```
[148] # Confusion matrix
      confusion_matrix = sklearn.metrics.confusion_matrix(Y_test, Y_pred)
      print(confusion_matrix)
      print("Correct Atheist (TN): " + str(confusion_matrix[0][0]))
      print("Correct Christian (TP): " + str(confusion_matrix[1][1]))
      print("Incorrect Atheist (FN): " + str(confusion_matrix[1][0]))
      print("Incorrect Christian (FP): " + str(confusion_matrix[0][1]))
```

```
[[276  43]
 [  4 394]]
Correct Atheist (TN): 276
Correct Christian (TP): 394
Incorrect Atheist (FN): 4
Incorrect Christian (FP): 43
```





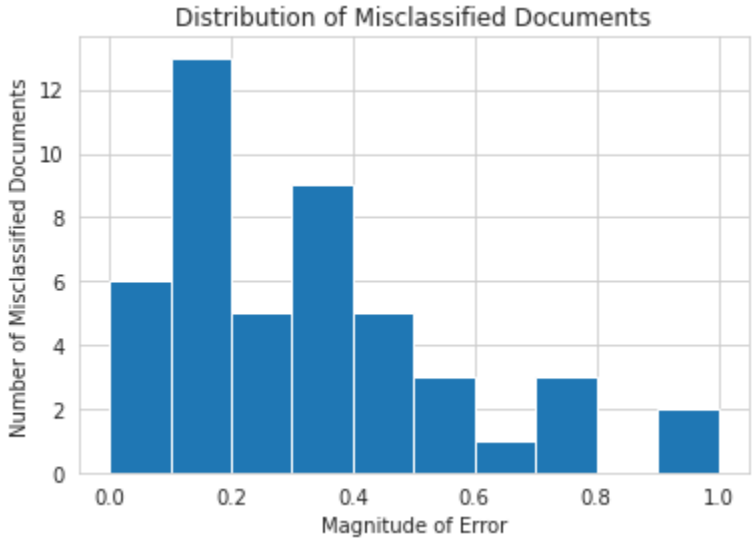
(c) (20 points) Use SHAP's explainer to study misclassified documents, and the features (words) that contributed to their misclassification, by taking the following steps:

- Report the accuracy of the classifier, as well as the number of misclassified documents.

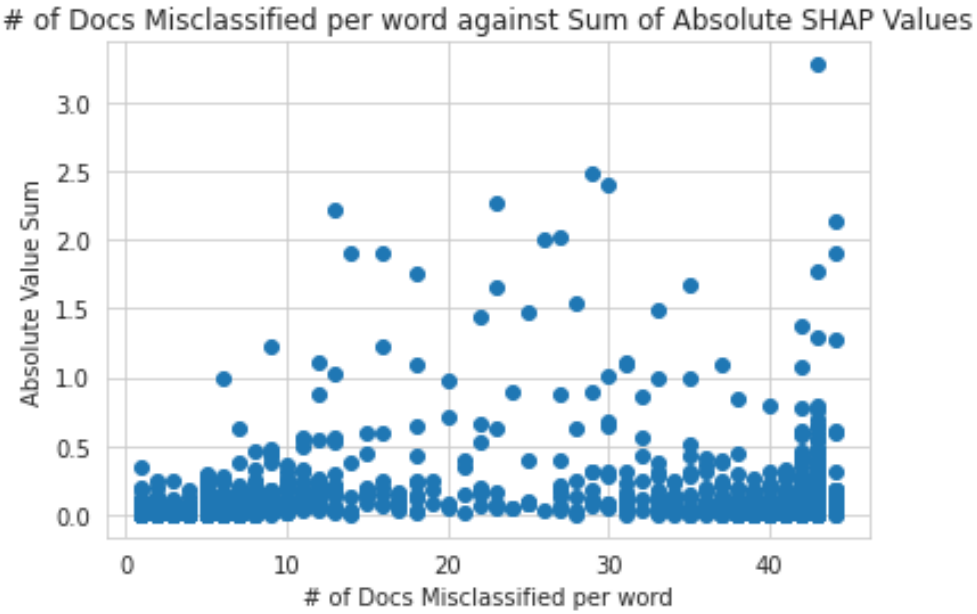
```
[141] # Compute the accuracy of the classifier and the number of misclassified documents
      print(f"Accuracy: {sklearn.metrics.accuracy_score(Y_pred, Y_test)}")
      print(f"Misclassified Docs: {np.count_nonzero(Y_pred != Y_test)}")

      Accuracy: 0.9344490934449093
      Misclassified Docs: 47
```

The accuracy of the classifier is 0.93, and there were 47 misclassified documents.

- For a document doc_i let us denote by conf_i the difference between the probabilities of the two predicted classes for that document. Generate a plot that shows conf_i for all misclassified documents (which, for misclassified documents, represents the magnitude of the error). Use any chart type you find appropriate to give a good sense of the distribution of errors.



- Identify all words that contributed to the misclassification of documents. (Naturally, some words will be implicated for multiple documents.) For each word (call it word_j), compute (a) the number of documents it helped misclassify (call is count_j) and (b) the total weight of that word in all documents it helped misclassify (weight_j) (sum of absolute values of weight_j for each misclassified document). The reason to use absolute values is that SHAP assigns a positive or a negative sign to weight_j depending on the class to which word_j is contributing. Plot the distribution of count_j and weight_j, and discuss your observations in the report.



The scatter plot shows a high concentration of absolute value sums between 0 and 0.5. This shows that the weight of most words that misclassified the documents was not that high. The plots are also concentrated

towards the right and left of the graph, between the 0-10 and 30-40+ range. This means that even though the weights of most words are low, the high frequency of the words at the low weight could be what led to the misclassification. The concentration of points between the 0-10 range means that many words have a low number of documents misclassified by them. It is interesting to note that the average weight in this segment appears to be lower, as we can see the maximum point in this segment is significantly lower than in the other ranges. This could also indicate that the lower average weight leads to a lower number of documents classified. There is also the possibility that the words in the 0-10 range are usually used in different contexts, hence why they ended up misclassifying the document because they were being used in a context that they weren't typically used in. On the other hand, the words in the 30+ range seem to have a higher average weight, which makes sense because it would contribute to a higher number of documents being misclassified. This could also be because the words that misclassify certain documents probably misclassify other documents too.