Emily Do, Naila Hajiyeva, Jacob Salazar, and Ethan Wen

For our final project, our team chose to study the presence of bias in advertising, particularly how it is used in converting their target audience, with the conversion being defined as users interacting with the ad. We wanted to examine how advertisements are targeted for certain users based on user information that could have been collected from them, such as age and gender. As such, we were interested to see if any particular features about users are more strongly considered in predicting user conversion. Before we conducted this feature importance analysis, we had expected to discover an existing bias in specific features that can be easily learned about users and would offer more user context, particularly the age and gender of the user. We hypothesized that advertisers may believe that users can be easily and successfully targeted based on the implications of their age and gender, pointing to the significance of those features.

We selected the *RandomForestRegressorregressor* from *sklearn.ensemble* because of the relatively short time it takes to run, as well as the ease of usability. We also used *NumPy*'s *np.random.seed()* to make sure our report is reproducible and unbiased. Another reason for using it was to not overtrain the model by assigning a constant to the *random_state*, potentially prompting the Regressor to memorize the inputs and outputs, thus delegitimizing our results. We also used the split of 1:3 instead of 1:4 since that is the amount that is predicted to be most efficient for a dataset with 8 parameters.[1] Additionally, we selected feature_importance_ to return the probability of each parameter influencing the outcome of the Regressor. We plotted the importance, to discover that the most influential features were income bracket, gender, and homeowner status.

We are working with the Synthetic dataset on Bias in Advertising Data that has been released by the IBM Developer team consisting of Ketan Barve, Karthikeyan Natesan Ramamurthy, Josh Price, Vishnupriya Pradeep, Skyler Speakman on June 15, 2022. We selected this dataset because it is very well-detailed and tests our knowledge of several skills, including testing for Null values, encoding categorical variables, rewriting columns for the code to run more smoothly, and ranking the features in the order of importance. After cleaning the dataset's age variable from the unknowns, we were left with 2390 values. Using this table, we used the confusion matrix in *sklearn.metrics* to check for true positive, false positive, false negative, and true negative values between the *true_conversion* and *predicted_conversion*. Separating another table that only consists of false negatives and false positives, we get 8 values total. For data visualization purposes, we used *seaborn*'s *kdeplot* together with *matplotlib* to plot density estimate plots of predicted conversions.

In the data analysis, we learned the most influential features in the dataset, and how they could be creating bias. Additionally, we found that by using *value_counts()* and plotting the distributions the majority of the dataset rows consisted of people between the ages of 55 and 64, the majority of college-educated, and the majority of homeowners. However, gender has split with twice as many females as males. The income label was split nearly evenly, with a few more people making below $100k than above. Additionally, it is important to note that the root means squared error rate was around 0.04 and the Mean Absolute Error was 0.02. The R^2 coefficient was approximately -0.03, implying there is no significant correlation in the prediction.

As we all know, predictive modeling is quite prevalent in the modern world; from the music on your phone to the recommendations on your Netflix account, and of course, to the

advertisements you see on your when watching Youtube or scrolling through social media. As stated before, our study used predictive modeling to determine if there was any *bias* present in the decision-making process of showing certain ads to people, based on a variety of factors. After an analysis of our findings, it is clear that they can be generalizable to other contexts, and for similar research purposes. To start, the use of “feature importance” in our study was one of the most important indicators to us of the presence or absence of bias in our dataset, and in the model's predictive capabilities. Because of this, we believe the use of feature importance can be utilized in a variety of situations that use predictive modeling (and not just to detect bias). In terms of the Telco churn model, feature importance can be used to see which factors most influenced the people who *did not* churn, and vice versa, to see which factors most influenced the people who *did* churn. Feature importance may also be used in a model predicting the efficiency of safety features in a new car; maybe the use of a certain locking mechanism in a seatbelt is “more important” to the car's safety than the thickness of glass used in the windows. In general, the feature importance’s usefulness and criticality in machine learning are evident. It allows for models to be tweaked and modified, ever so slightly (or drastically in some cases), to improve and/or upgrade said model.

As mentioned before, after subsetting the original dataset of about 1.5 million data points, down to 2390, we found that the predictive model had high accuracy in predicting whether or not a person (based on their gender, income, age, and homeowner status) clicked on the ad. *This indicates that there was relatively no bias in the model's predictions*. For transparency purposes, the drastic reduction in size from the original dataset may have skewed the results of our study, but the model was not perfect, showing bias in a particular variable, and therefore requires a solution to improve model accuracy. An issue arises when comparing the outliers of the model to other data points in the model that were deemed to have a correctly predicted and true conversion. In total three data points were predicted to convert (click on the ad) while five were predicted to not convert (ignore the ad). Out of the eight outliers, all were college educated and between the ages of 55-64. This subset of outliers was split evenly between males and females but differed in their *income* and *homeowner* status. When analyzing the other 2381 data points, most were given compatible predicted and true conversion values (i.e. the model correctly predicted whether or not they clicked on the ad), but more specifically, data points that *were* homeowners were correctly predicted to *not* click on the ad. Coupling the analysis of the outliers and correctly predicted data points, a clear bias is present in the homeowner variable, with a slight bias in the income variable as well. Thanks to feature importance, we can be confident in that statement since “income” and “homeowner” were more “important” to the model's prediction than other variables.

To mitigate the bias present in the dataset, specifically in the income and homeowner variables, we propose that the meaning of the“homeowner” variable be changed from owning or not owning a home, to the residence type of each person (house, apartment, college dorm, co-op, etc.). We also propose the “income” variable be completely taken out of the dataset. We chose these solutions because changing the meaning of the “homeowner” variable would result in a more varied dataset since it would no longer rely on the binary “yes” or “no”, and instead would utilize a numerical system that represented the type of residence of each data point. Also, the model would no longer target homeowners and non-homeowners, and would instead have more to choose from, decreasing the chance of bias in its predictions. Removing the “income”

variable from the dataset enables other variables to be more impactful in the model's predictions. Simply modifying the income variable wouldn't make much of a difference in the data since our sample size is almost evenly split between <100K and >100K, as evidenced by the income distribution chart. We believe our proposed solutions are fair because they decrease any discrimination against prospective data points (i.e. people), and instead allow for a more inclusive dataset.

In closing, our model concerning targeted advertising, through predictive modeling that uses feature importance, found bias in two variables, "homeowner" and "income". Although a majority of the model's predictions were correct (true conversion), the outliers showed clear bias in its scores for non-homeowners making less than 100K. Changes to the dataset to make it more accurate, and more importantly with less bias, include altering the "homeowner" variable to include residence type, and removing the "income" variable altogether. These solutions may result in a more inclusive dataset, with little to no bias. We were correct in our hypothesis that some features/variables would be the cause of bias in the model, though were incorrect in hypothesizing the most impactful ones next to gender. The "income" and "homeowner" variables were just as impactful on the models' predictive scores, with "income" being the most "important". Advertisers must assess the ethical implications of their predictive models, seeing as bias can be prevalent in a model's processes.

Reference:

1. Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *15*(4), 531–538. doi:10.1002/sam.11583