Who is my target audience?

Lindsey Vanosky • 08.04.2022

Overview

Stakeholder:

Amy White, Marketing Manager

Business Problem:

Amy is tasked with identifying keywords and demographic information to use as part of a paid marketing push for an online class to improve your career.

She wants to target individuals making less than $50K a year.

Selecting a Data Set

US Census Data

- For this project, I chose a data set from Kaggle that came from the US Census. This data set has the following set of features:

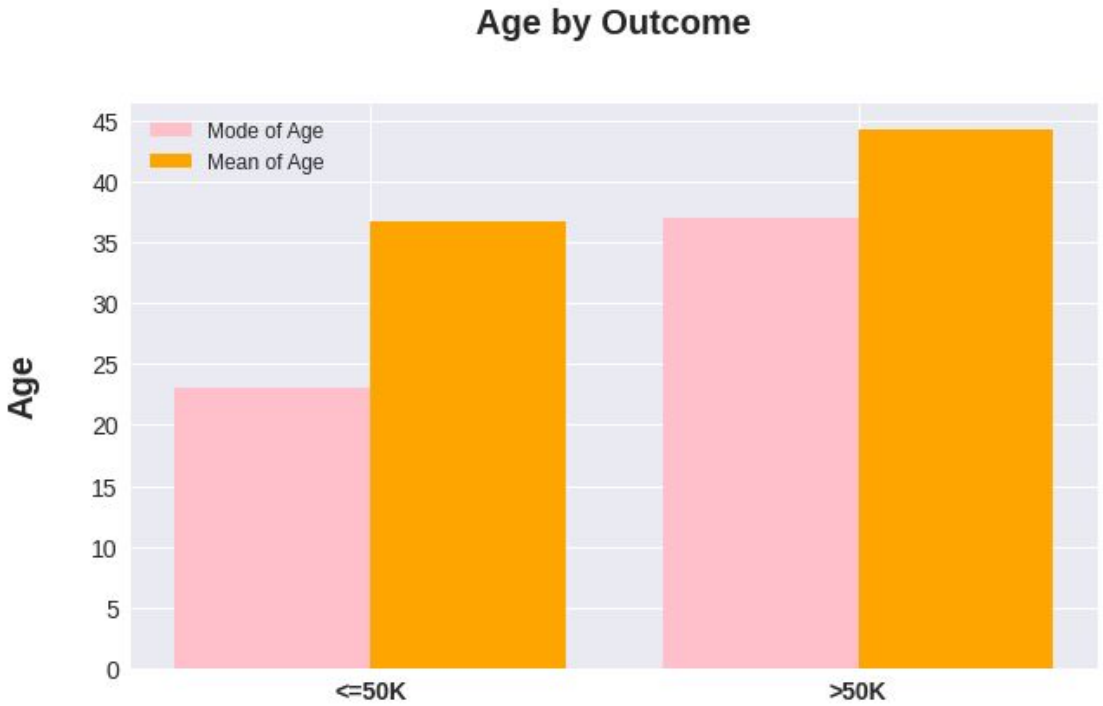
- The original data set had 32,561 entries and 15 columns. The data set used here had 32,537 entries and 13 columns after dropping irrelevant information.

```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32537 non-null  int64
 1   workclass       32537 non-null  object
 2   education       32537 non-null  object
 3   marital_status  32537 non-null  object
 4   occupation      32537 non-null  object
 5   relationship    32537 non-null  object
 6   race            32537 non-null  object
 7   gender          32537 non-null  object
 8   capital_gain    32537 non-null  int64
 9   capital_loss    32537 non-null  int64
 10  hours_per_week  32537 non-null  int64
 11  native_country  32537 non-null  object
 12  outcome         32537 non-null  object
```

Visual 1 - Line Graph of Salaries by Race

Given that our business problem is to identify patterns in data to advise keywords for a digital marketing campaign, I wanted to dive into age. Age is an important factor when setting campaign parameters. If we were to just look at the mean, we'd be overshooting our target audience because the bulk of those who make under $50K are 23.

Using a range of 23 - 45 would provide better results.

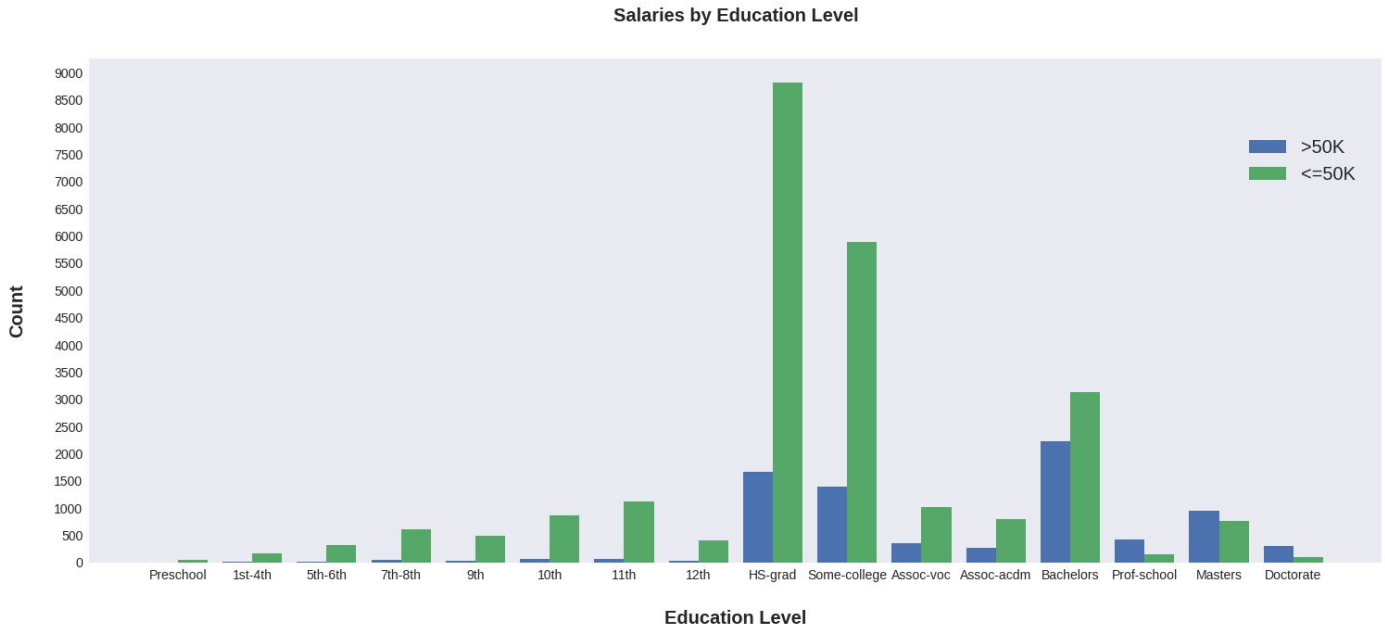


Visual 2 - Bar Chart of Salaries by Education

This bar chart represents salaries by education level. As we can clearly see it pays to stay in school! The highest concentration of salaries >$50K is from HS grad and beyond.

Prof school, masters, and doctorate, are the only columns where those earning >$50k outnumber those making <=$50K.

Meaning Bachelors degree and under will make the best keywords.

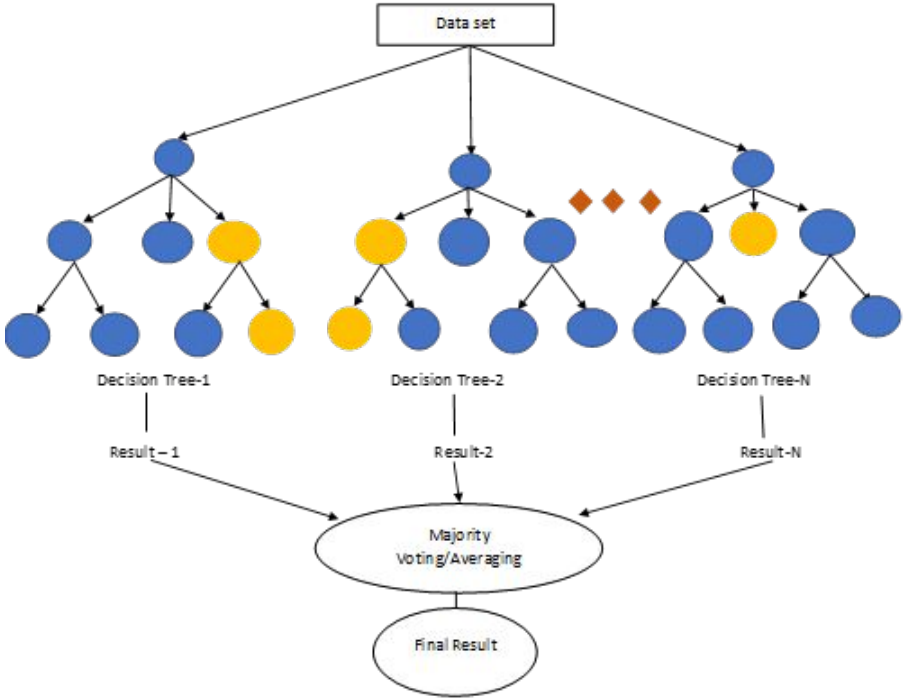


This visual gives us a good idea of what keywords would be useful.

Machine Learning

Model - Random Forest Classifier

Using the insights gathered from the explanatory visuals, it was decided to build a Random Forest Classifier algorithm to find patterns throughout the data and predict whether or not someone makes <=$50K based off of demographic features.

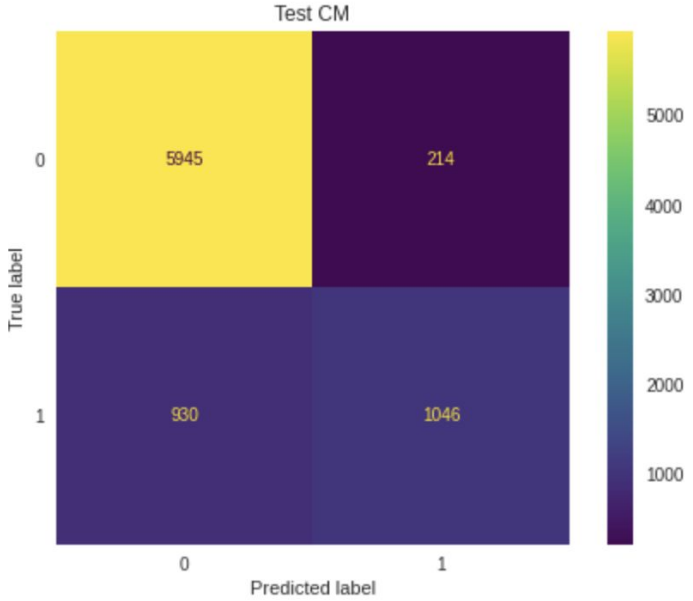


Performance Metrics

One strength of this model is that is scores highly in precision and accuracy on the <=$50K outcome set. Since this is the audience we are targeting, the performance on the >$50K outcome set isn't as important.

It was with this model we were able to achieve our highest number of True Negatives. A true negative in this case means that we predicted that they would make less than $50K, and in fact they do.

In summary, our model is 86% accurate when it comes to predicting if someone makes less than $50K.



Summary

1. This model can successfully predict whether someone makes less than $50K based off of demographic information.

2. Using a combination of education, workclass, and occupation features as keywords will result in the target audience.

3. Be cautious when using demographic data to ensure results are equitous.