Where is the Optimal Location for my Boutique?

Lindsey Vanosky • 08.04.2022

Overview

Stakeholder:

Amy White, Commercial Developer

Business Problem:

Amy is tasked with finding the best location for a new high end boutique. She has scouted three possible locations and is looking to see which is most optimal for her client base.

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:

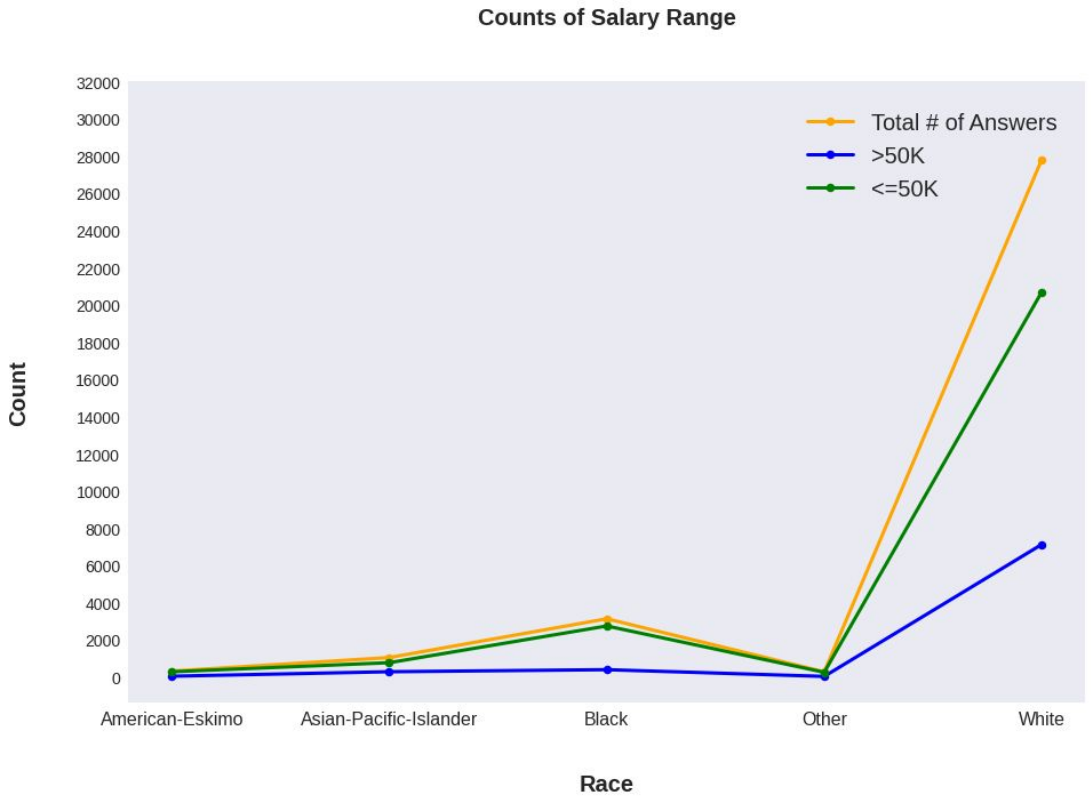
- The original data set had 32561 entries and 15 columns. The data set used here had 32537 entries and 13 columns after dropping irrelevant information.

```
 #   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

This line plot represents salaries by race. This census data categorized salaries as either >50K or <=50K. At first plot, it looked like there was a massive disparity of >50K by race. It looked as if salaries over 50K were disproportionately awarded to white people. This isn't necessarily untrue, but I found it important to plot the line of total answers by race so we can see who this data actually represents. Adding this line did highlight that the majority of our data was from white people. Therefore my first recommendation - more data!

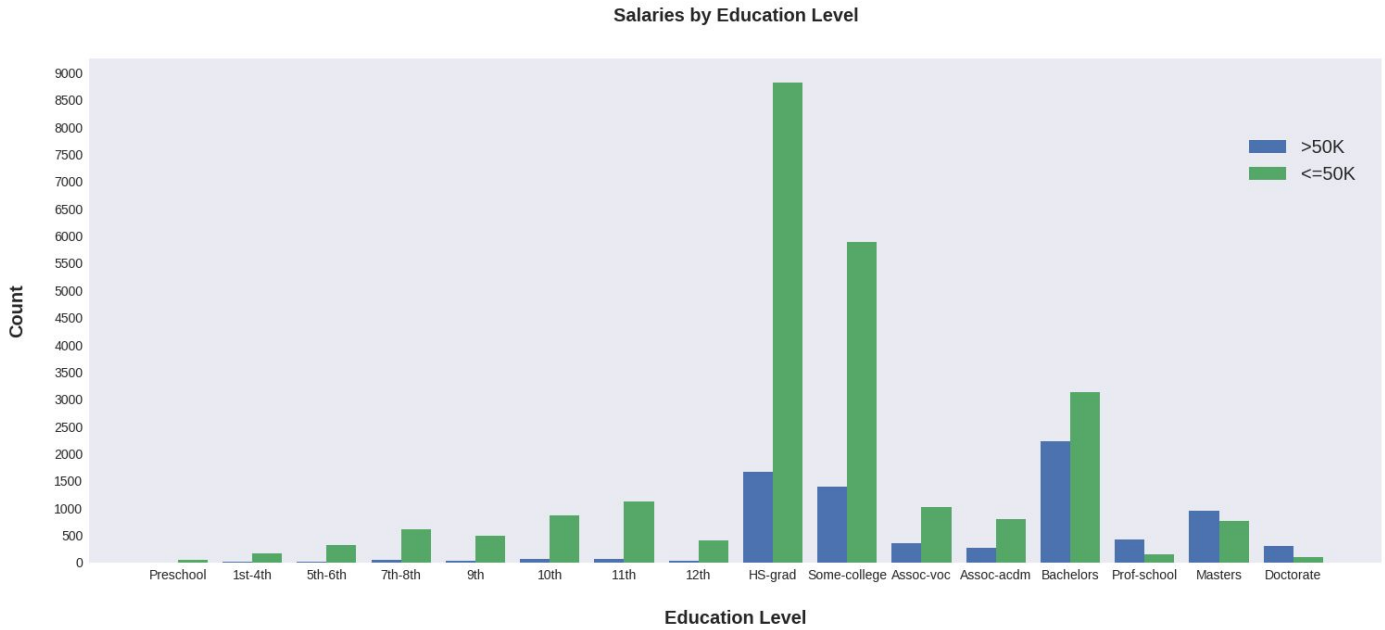
However, plotting the total answers line also showed that there was indeed a disparity primarily in the Black category. We can see there is a significant gap between the orange line and the blue line in this category, indicating that the overwhelming majority of black people represented here make less than $50K.



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.

What's interesting is that there is a large disparity in number of people making >50K amd <=50K in the HS and some college columns. I'd recommend more data and if we still see the same thing, it'd be worthwhile to examine the other features of those that fall in each of those categories. What makes one HS grad make more than another?



The fact that the total number of >50K is very similar between the HS grad and Bachelors columns, makes me think there could be opportunity to use this data set as a tool to decide whether to go to college or not base on finances.

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

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.

One strength of this model is that is scores highly in precision and accuracy on the <=$50K outcome set. This model is also precise on the >$50K outcome set, however, we get a low recall score. This means we are likely to see relatively more false negatives.

In this case, a false negative would mean that the individual was predicted to make less than $50K and instead does make more than $50K. This incorrect prediction is incredibly low risk, so it is more optimal to have a model that has more false negatives than false positives. Which is what we have achieved here.

Summary

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

2. More data for the >$50K outcome set is recommended. We are currently working with an unbalanced data set.

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