UBC Salaries: Exploratory Analysis of Gender

Jade Bouchard

Table of contents

Aim

This document explores The University of British Columbia (UBC) faculty salaries based on guessed gender.

Data

Salary data was sourced from The University of British Columbia (2020). To access individual financial reports, click on the yearly links under the header `Statement of Financial Information (SOFI)`.

Gender data was inferred using first names of staff members. In order to guess gender, I used baby name datasets (Statistics Canada, n.d.; Kaggle 2017; Sharma 2020).

Ethics

In this project first names are used to guess whether someone is "male" or "female". Many people do not fit into these categories. Misgendering, or incorrectly assigning gender to individuals, can have harmful effects. In addition, while first names can sometimes be an indication of someones gender, first names are not inherintly gendered.

I encourage anyone who notices a misgendering within this project to raise an issue in the issues tab, and it will be corrected. In addition, I encourage respectful and inclusive language in all discussions related to gender.

Methods

The Python programming language (Van Rossum and Drake 2009) was used to perform this analysis.

Data collection

As mentioned earlier, for UBC salary data, I used the salary PDFs that UBC releases every year (The University of British Columbia 2020). The following steps were taken to collect the data.

- Use the `requests` package to access the UBC Financial Reports webpage.
- Extract all links and filter for the Statement of Financial Information (SOFI) PDF links which contain salary information.
- If there are any PDFs from which I have not already collected salary data, extract the text from the PDF using the package `pypdf`.
- Store the new salary data

An exerpt of the raw salary data is below.

```
Kolhatkar, Ashra  79,036   550  Kolhatkar, Varada  88,579   -  Kolind, Shannon H  108,817
```

Data cleaning

In this section, the following steps are taken to clean the salary data:

For each year of salary data,

- Remove special charachters from text. For example, ş.
- Remove unnecessary text content. For example, the "[Auditor's] Qualified Opinion".

- Uses regex to process the raw text data into a structured DataFrame with columns: `Name`, `Remuneration`, `Expenses`.
- Split the `Name` column into first and last names.
- Convert salary (remuneration) and expenses columns to a numeric data type.
- Shorten first and last names to allow for easier name matching between years. For example, someone's name in 2020 could be "Bob M Sherbert" and in 2021 their name could be "Bob-M Sherbert". This name would be shortened to Bob Sherbert to avoid mismatching.

Then, I concatenate dataframes from all years together.

Table 1 shows an expert of the cleaned salary data.

Table 1: Clean UBC Salary Data

	Last_Name	First_Name	Remuneration	Expenses	Year
0	Aamodt	Tor	193153	5597.0	2023
1	Abanto	Arleni	107723	393.0	2023
2	Abbassi	Arash	109136	82.0	2023
3	Abdalkhani	Arman	101829	NaN	2023
4	Abdi	Ali	238203	2981.0	2023

Gender Prediction

Babyname Corpus

In order to predict gender, I used datasets with babynames and assigned genders. In order to have a somewhat diverse set of baby names, I used babynames from Canadian, American, and Indian sources Sharma (2020).

For each UBC staff name, I found whether that name was more common among girls or boys in the babyname corpus. Then, I guessed the gender that was most common.

In Table 2, the `Estimated_Accuracy` column shows the percentage of gender majority from the babyname corpus. For example, if 95% of babys named George in the dataset were male, the `Estimated_Accuracy` column value would be 0.95. Any staff names that had less than an 80% gender majority were given a null gender.

Table 2: Babyname Data

	index	Sex_at_birth	First_Name	Estimated_Accuracy
95660	24125.0	Female	Irin	0.5

Table 2 shows the name Irin has about the same number of boy and girl names. This name has an accuracy of 0.5. So, since its not above the 80% threshold, anyone with that name would have a gender value of `None` assigned.

Table 3 shows some of the predictions made on UBC staff members using the babyname corpus.

Table 3: Babyname Data

	First_Name	Sex_at_birth	Estimated_Accuracy
0	Tor	Male	1.0
1	Arleni	Female	1.0
2	Arash	Male	1.0
3	Arman	Male	1.0
5	Yasmine	Female	1.0

Around 91.98% of UBC staff names were matched with names in the babyname corpus. 96.2% of the corpus matches had over 0.8 accuracy and were kept, the rest were given a prediction of `None`.

NLTK

For UBC staff names that were not found in the babyname corpus, I used a natural language processing model to predict genders (Bird, Loper, and Klein 2009).

Table 4 shows some examples of names not found in the babyname corpus.

Table 4: Example Staff Names Not Found in Babyname Datasets

	First_Name
0	Fatawu
1	Tamiza
2	Purang
3	Ninan
4	Reto

In order to train and evaluate the model, the babyname corpus was split into a trianing and test set.

Features used to train the Naive Bayes model were the last 2, 3, and 4 letters of the staff member's first name.

The accuracy on the test set was 0.86. However, the accuracy on the UBC staff names that were missing from the babyname corpus is very likely lower than 0.86. This is due to the fact that the data we are making predictions on is quite different from our training data.

Below are the top three features the classifier found most useful for making correct predictions.

```
Most Informative Features
          last_4_letters = 'isha'         Female : Male   =    305.2 : 1.0
          last_4_letters = 'etta'         Female : Male   =    193.7 : 1.0
          last_3_letters = 'cia'          Female : Male   =    179.7 : 1.0
```

We can see there are patterns in first names that could be helpful for predicting gender. However, these patterns may not show up often in the unique UBC staff that were not in the babyname datasets.

Finally, after making predictions on the UBC staff data, we can see in Table 5 the predictions our classifier was least confident about, and in Table 6 and the predictions it was most confident about.

Table 5: Least Confident NLTK predictions

	First_Name	Sex_at_birth	Estimated_Accuracy
1097	Faride	Female	0.43
694	Jiahua	Female	0.44
1646	Xiaohua	Female	0.44
1404	Chuan-Hui	Male	0.44
1092	Qingshi	Female	0.44

Table 6: Most Confident NLTK predictions

	First_Name	Sex_at_birth	Estimated_Accuracy
1501	Sav	Male	0.86
1183	Ninan	Male	0.86
1181	Tamiza	Female	0.86
1179	Bingshuang	Male	0.86
2213	Cicie	Female	0.86

To represent the extra uncertainty in using a classifier compared to the babyname corpus, the `Estimated_Accuracy` column for NLTK predictions is the classifier's predict-proba score multplied by 0.86. Where 0.86 is the accuracy of the classifier on the test set.

Like with the babyname corpus predictions, NLTK predictions with an accuracy less than 0.8 were given a gender prediction of `None`. This was the case if the predict-proba score from the classifier was less than 0.93.

Overall, 80.2% of the NLTK predictions were kept and the rest were given a prediction of `None`.

Exploratory Data Analysis

Limitations

Due to the low level of accuracy in gender predicion of this report, conlusions drawn are not meaningful.

- NLTK training data not representative of real data
- no practical way to check ground truths
- data sources
- gender not binary
- Naive Bayes conditional independence assumption violated These features are correlated which violates the Naive Bayes conditional independence assumption. However, I decided to stick with the NLTK naive bayes classifier as it efficiently handles text data, large datasets, and still performs fairly well.
- etc.

References

Bird, Steven, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python.*

Kaggle. 2017. "US Baby Names." https://www.kaggle.com/datasets/kaggle/us-baby-names/data.

Sharma, Anany. 2020. "Indian Names Dataset." https://www.kaggle.com/datasets/ananysharma/indian-names-dataset.

Statistics Canada. n.d. "Table 17-10-0147-01 First Names at Birth by Sex at Birth, Selected Indicators (Number)." https://doi.org/https://doi.org/10.25318/1710014701-eng.

The University of British Columbia. 2020. "UBC Statement of Financial Information." Open Government - Open Data. https://finance.ubc.ca/reporting-planning-analysis/reports-and-disclosures.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace.