US BABY NAMES PROJECT

BY SIMRAN CELLINI

PRESENTATION OUTLINE

- Project motivation
- Project aim.
- Gathering and analyzing my data.
- Approach.
- Data Exploration.
- Data Exploration Part 2.
- Methodology.
- Naïve Bayes Result
- Decision Tree Results.
- Support Vector Machine Results.
- Conclusion

PROJECT MOTIVATION

When researching into this topic I found out the following:

- 57% of American parents said that girls were easier to name than boys.
- Parents over 35 years old are more likely to choose a classic or old-fashioned name.
- Younger parents, on the other hand, are more attracted to uncommon or unique names.

After looking at the data online I noticed that there were similar trends in how babies born in the US are named.

PROJECT AIM

BASED ON THE NAME OF A BABY CAN WE PREDICT WHETHER THEY ARE MALE OR FEMALE?

GATHERING & ANALYSING MY DATA

The first step was to gather and analyze my data. The table shows a list of baby names, the year they're born starting from 1910 – 2014, whether the gender of the baby was Male or Female. The table has 50.000 baby names in it.

Id	Name	Year	Gender	State	Count
1	Mary	1910	F	AK	14
2	Annie	1910	F	AK	12
3	Anna	1910	F	AK	10
4	Margaret	1910	F	AK	8
5	Helen	1910	F	AK	7
6	Elsie	1910	F	AK	6
7	Lucy	1910	F	AK	6
8	Dorothy	1910	F	AK	5
9	Mary	1911	F	AK	12
10	Margaret	1911	F	AK	7
11	Ruth	1911	F	AK	7
12	Annie	1911	F	AK	6
13	Elizabeth	1911	F	AK	6
14	Helen	1911	F	AK	6
15	Mary	1912	F	AK	9

ID: Number of baby names in the table.
Name: Different types of baby names.
Year: Baby names from 1910 – 2014.
State: Different states in America. 10 states out 50 states.
Gender: Whether the name is Male or Female.
Count: The number of baby names in that particular year in that state. For example there are 14 babies were named Mary in the year of 1910.

Important features from the table:

- The first letter.
- The second letter.
- The last letter.

APPROACH

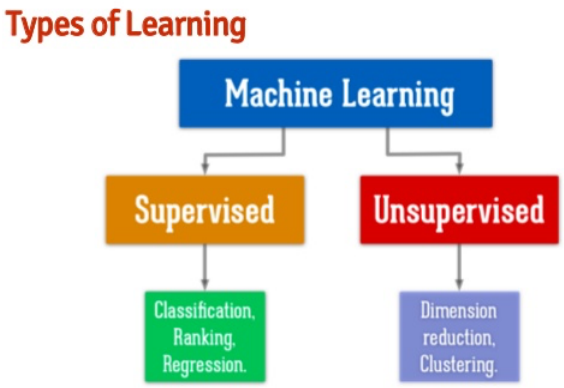
Inputs were taken from the table.

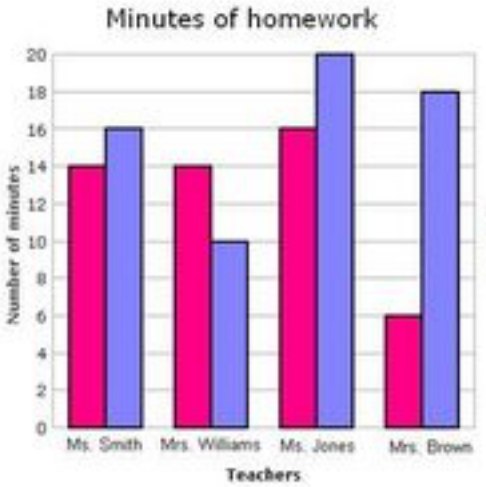
Model built on the training set to predict whether the gender is male or female for new data.

Validate / Test the model on the test set . Explore data to understand which inputs are most important.

Analyse the results.

Id	Name	Year	Gender	State	Count
1	Mary	1910	F	AK	14
2	Annie	1910	F	AK	12
3	Anna	1910	F	AK	10
4	Margaret	1910	F	AK	8
5	Helen	1910	F	AK	7





DATA EXPLORATION

- I found inputs that were good predictors of gender such as the first letter, the second letter and the last letter and added them into the table in separate columns as below.

	Id	Name	Year	Gender	State	Count	last_letter	first_letter	second_letter
932175	932176	Monica	1989	M	DC	16	a	M	o
434483	434484	Rita	1978	F	CA	135	a	R	i
984289	984290	Melody	1947	F	FL	14	y	M	e
547037	547038	Lilyanne	2010	F	CA	16	e	L	i
816512	816513	Sandra	1938	F	CT	159	a	S	a

The table shows baby names from 1910 to 2014.

- The two tables below were good predictors of genders as it showed me that the last letter and the first letter had a huge impact in predicting whether a name was male or female.

The last letter table.

Gender	count	unique	top	freq
F	29802	23	a	11009
M	20198	27	n	4833

The first letter table.

Gender	count	unique	top	freq
F	29802	26	A	3093
M	20198	26	J	2035

DATA EXPLORATION PART 2

- During my data exploration I also wanted to find out what the most popular names were during specific decades. I created word clouds. This word cloud shows the most popular names from the 1920s – 1930's. The most popular names were David and Richard.

I also created other word clouds for the other decades. By creating multiple word clouds I can compare what the most popular baby names were from different decades.

METHODOLOGY

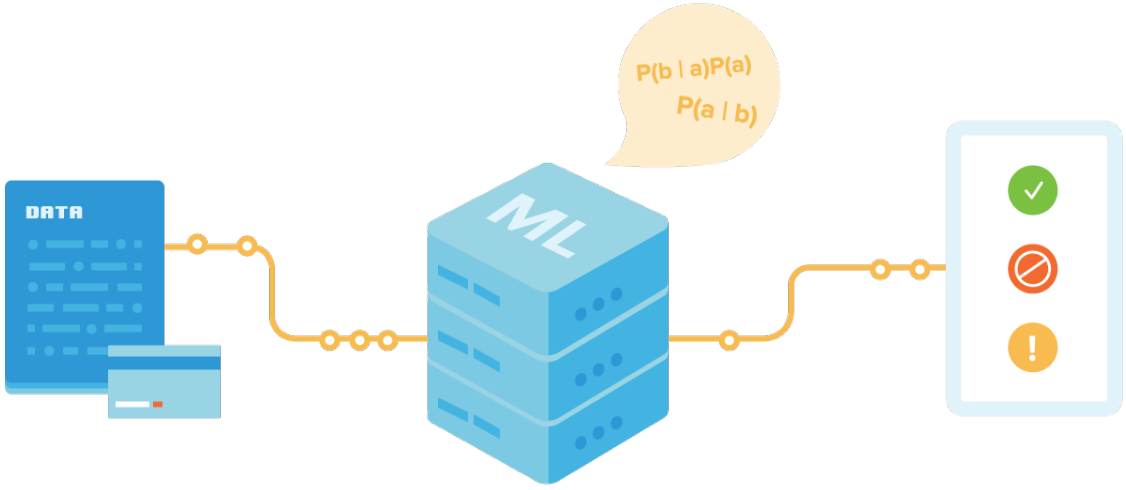
Below are the three machine learning models that I used.

Models

Naïve Bayes (Looks at each feature on its own, looks at the relationship)
Decision Trees (Makes predictions by choosing one input at a time splitting the data)
Support Vector Machine (Splits the two classes by the inputs, the largest separation the better)

Evaluation

Confusion Matrix
Accuracy



NAÏVE BAYES RESULTS

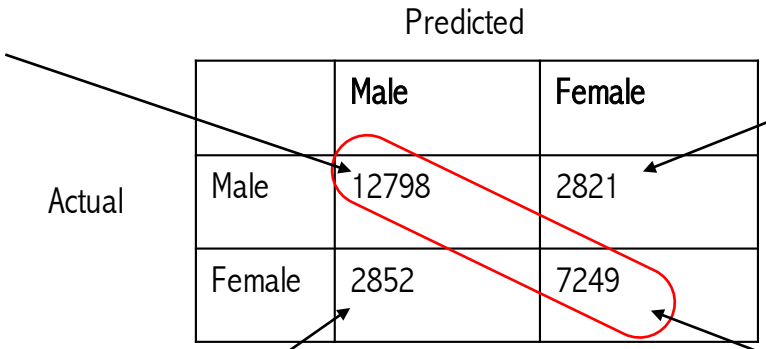
CONFUSION MATRIX

Predicted male names to actual male names 12798.

Predicted Female names to actual Male names 2821.

Predicted male names to actual female names 2852.

Predicted Female names to actual female names 7249.



ACCURACY

77 %

DECISION TREE RESULTS

CONFUSION MATRIX

Predicted male names to actual male names 13139.

Predicted

Predicted Female names to actual Male names 1760.

	Male	Female
Male	13139	1760
Female	2238	7863

Actual

Predicted male names to actual female names 2238.

ACCURACY

Predicted Female names to actual female names 7863.

84 %

SUPPORT VECTOR MACHINE RESULTS

CONFUSION MATRIX

Predicted male names to actual male names 12089.

Predicted

Predicted Female names to actual Male names 2839.

	Male	Female
Male	12089	2839
Female	2810	7262

Actual

ACCURACY

Predicted male names to actual female names 2810.

77 %

Predicted Female names to actual female names 7262.

CONCLUSION

- The best model to use was the Decision Trees. It shows that by looking at certain letters in a name, we can predict whether the name is male or female. It also shows that the model predicted the correct gender of the name with 84 % accuracy.

- The last letter had an huge impact in the model predicting whether a name was male or female.
- Approximately 10000 names with the last letter "A" were Female.
- Approximately 5000 names with the last letter "N" were Male.

Gender	count	unique	top	freq
F	29802	23	a	11009
M	20198	27	n	4833

- The results also showed that approximately 3000 female names started with the letter "A".
- The results showed that approximately 2000 male names started with the letter "J".

Gender	count	unique	top	freq
F	29802	26	A	3093
M	20198	26	J	2035

ANY QUESTIONS ?