DATAPLAY

GENDER PREDICTION ANALYSIS
BY NAME

USING MS EXCEL

SHRADDHA SAHA

Contents

- Problem statement
- Data collection and preparation
- Working approach
- Result

Problem statement

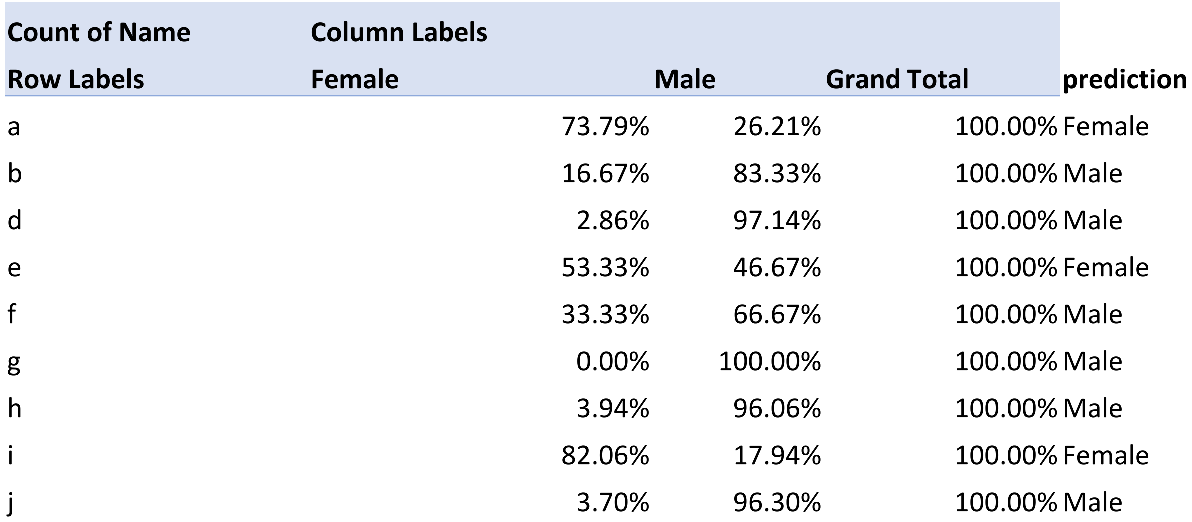
- A dataset with different names of females and males are given.
- With the last letter of the name, we have to predict the gender of the person.(Using Excel only)
- Predict gender using basic probability
- Calculate prediction's accuracy/f1 score

Data collection and preparation

- We have collected (spreadsheet provided by Dataplay) a dataset with a total of 3001 entries, where the names, last letter and actual gender are provided.
- To build our model, we have done a 70-30 train-validation split on the dataset.
- We have taken first 70% i.e, first 2101 entries for train data and last 30% i.e, last 900 entries for validation data.

Working Approach

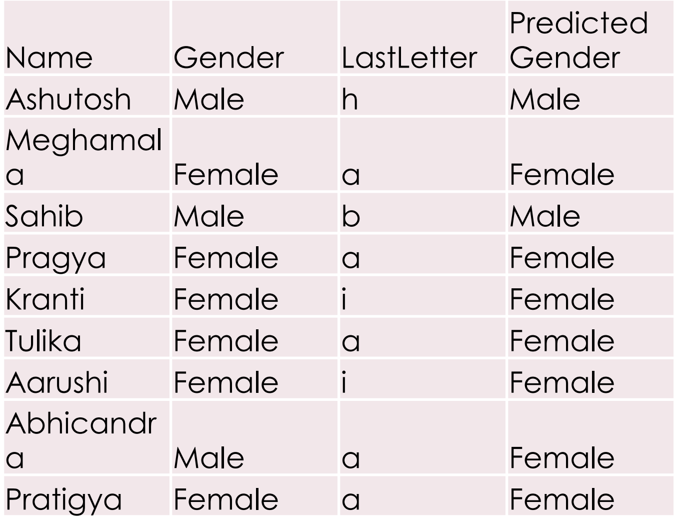
- To build our model, we used ‘Pivot table fields’ on train data.
- Obtained the male-female % by counting the last letter.
- For the prediction column, we used IF Statement on Female and Male columns.



Count of Name	Column Labels			
Row Labels	Female	Male	Grand Total	prediction
a	73.79%	26.21%	100.00%	Female
b	16.67%	83.33%	100.00%	Male
d	2.86%	97.14%	100.00%	Male
e	53.33%	46.67%	100.00%	Female
f	33.33%	66.67%	100.00%	Male
g	0.00%	100.00%	100.00%	Male
h	3.94%	96.06%	100.00%	Male
i	82.06%	17.94%	100.00%	Female
j	3.70%	96.30%	100.00%	Male

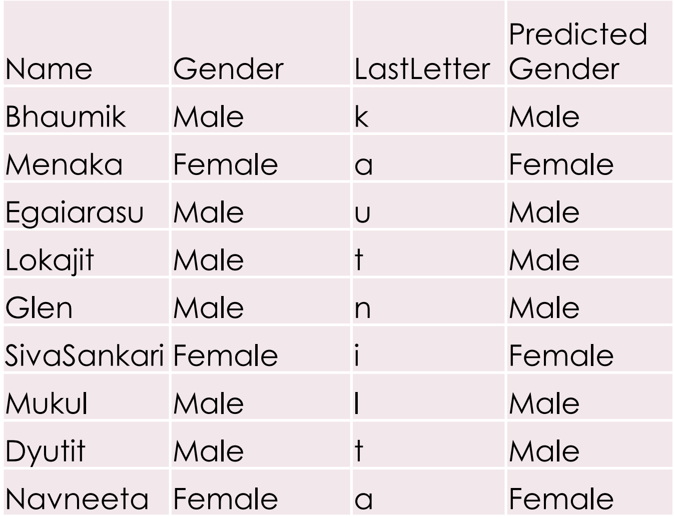
Working Approach (cont.)

- Next we used VLOOKUP formula to fit our model on train data and validation data.



Name	Gender	LastLetter	Predicted Gender
Ashutosh	Male	h	Male
Meghamala	Female	a	Female
Sahib	Male	b	Male
Pragya	Female	a	Female
Kranti	Female	i	Female
Tulika	Female	a	Female
Aarushi	Female	i	Female
Abhicandra	Male	a	Female
Pratigya	Female	a	Female

Train data

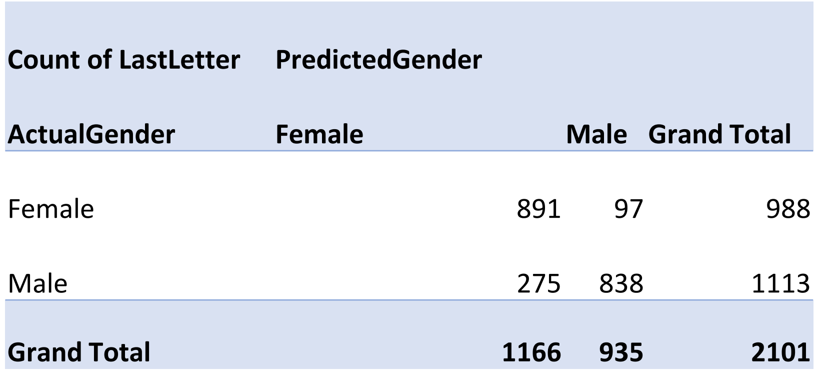


Name	Gender	LastLetter	Predicted Gender
Bhaumik	Male	k	Male
Menaka	Female	a	Female
Egaiarasu	Male	u	Male
Lokajit	Male	t	Male
Glen	Male	n	Male
SivaSankari	Female	i	Female
Mukul	Male	l	Male
Dyutit	Male	t	Male
Navneeta	Female	a	Female

Validation data

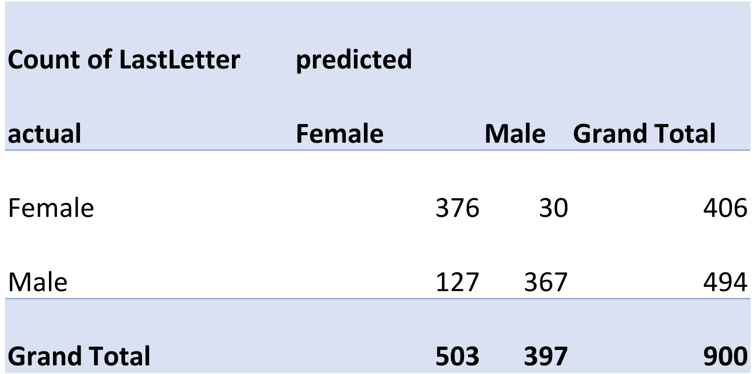
Working Approach (cont.)

- Now is the time to check if our model is good or bad.
- For this, we used PIVOT TABLE FIELDS to make the Confusion Matrix first.



Count of LastLetter	PredictedGender		
ActualGender	Female	Male	Grand Total
Female	891	97	988
Male	275	838	1113
Grand Total	**1166**	**935**	**2101**

Confusion matrix for train data



Count of LastLetter	predicted		
actual	Female	Male	Grand Total
Female	376	30	406
Male	127	367	494
Grand Total	**503**	**397**	**900**

Confusion matrix for validation data

Working Approach (cont.)

- First we calculated Accuracy on Train data.
- The accuracy metric computes how many times a model made a correct prediction across the entire dataset.
- Accuracy Score for Train data is **0.8229** and for validation data is **0.8255**
- This can be a reliable metric only if the dataset is class-balanced; that is, each class of the dataset has the same number of samples.
- Nevertheless, real-world datasets are heavily class-imbalanced, often making this metric unviable.

Working Approach (cont.)

- To overcome this problem, we use F1 Score which is a good indication of model performance.
- F1 Score is the harmonic mean for the class-wise precision and recall values.
- Precision measures how many of the “positive” predictions made by the model were correct.
- Recall measures how many of the positive class samples present in the dataset were correctly identified by the model.
- Here, we used positive=‘Female’ and negative=‘Male’ for calculation purpose. (other choice can also be done)

Working Approach (cont.)

- For Train Data, Precision Score(+ve)= 0.7641 , Recall Score(+ve)=0.9018
- So, F1 Score (positive) = 0.8272
- and Precision Score(-ve)= 0.8962 , Recall Score(-ve)=0.7529
- So, F1 Score (negative) = 0.8182

- Combining both the classes, we get Macro avg. F1 Score for train data.
- **Macro Average F1 Score(train data)** = Simple average of F1 scores of both the classes= (0.8272+0.8182)/2 = **0.8227**

Working Approach (cont.)

- For Validation data, Precision Score(+ve)=0.7475, Recall (+ve) = 0.9261
- So, F1 Score(positive)=0.8272
- And Precision Score(-ve)=0.9244, Recall (+ve) = 0.7429
- So, F1 Score(negative)=0.8237
- Thus, **Macro Average F1 score(Validation data)= 0.8254**

- We can see that the Macro Average F1 score is approximately same for both Train and Validation data; so, it seems there is no underfitting or overfitting problem.

Result

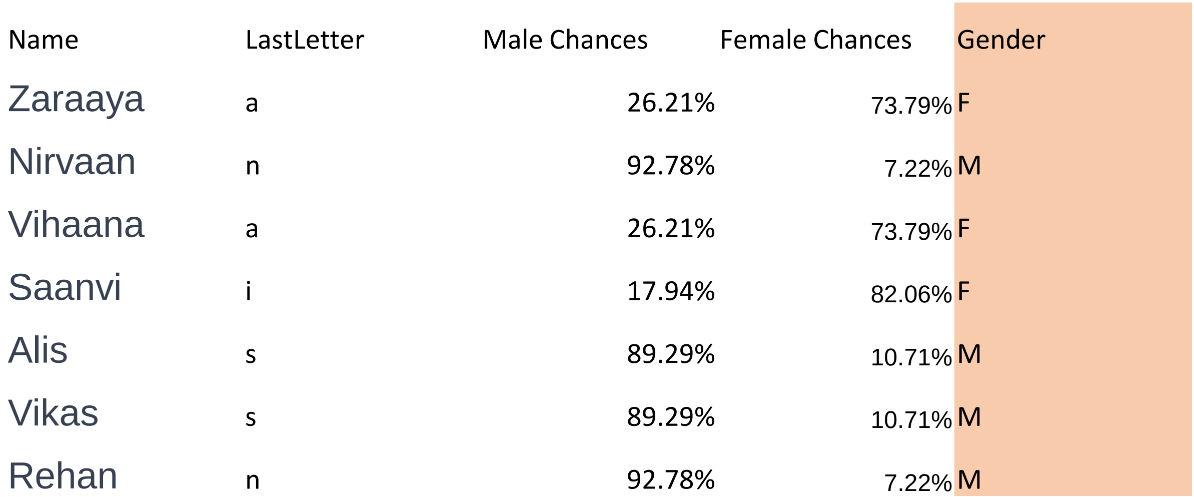
- Now, as we built our model on train data, it is obvious that model performance is good on train data. We will focus on Macro Avg. F1 Score on Validation data, it helps to understand how well our model generalizes to unseen data.

- From our obtained values, we have come to the conclusion that our model performance is good to predict the gender of a person from the last letter of their names as **Macro Average F1 score is quite high.**

- Now, we will take a test data of our choice to check how our model works.

Result

- It seems a good prediction on our test data.



Name	LastLetter	Male Chances	Female Chances	Gender
Zaraaya	a	26.21%	73.79%	F
Nirvaan	n	92.78%	7.22%	M
Vihaana	a	26.21%	73.79%	F
Saanvi	i	17.94%	82.06%	F
Alis	s	89.29%	10.71%	M
Vikas	s	89.29%	10.71%	M
Rehan	n	92.78%	7.22%	M

Test data

DATAPLAY

Thank you !