A Mathematical Essay on Naive Bayes

Aryan Pandey
Department of Ocean Engineering
Indian Institute of Technology, Madras
Chennai, India
na19b030@smail.iitm.ac.in

***Abstract*—In this study, we will study the mathematical aspects of the Naive Bayes Algorithm. The dataset that we use is the Census Dataset from the United States of America in the year 1994. Using Visualisation and Modelling techniques, the conclusions we draw from this study are three-fold. First, We see that there is a significant difference in behaviour of people who invest their money versus the ones who don't. Second, Some features like the age of the person and the number of years of education have a significant impact on the classification problem at hand. Third, dealing with the investors and non-investors separately improves performance.**

***Index Terms*—Naive Bayes, Classification**

I. Introduction

One of the key factors to look into when understanding the demography of a country, is to look at the income levels of the people residing in it. In many cases the simplest and fastest way to get an idea of the overall income levels of a country is to look at the number of people who earn above a certain threshold and what are the major traits associated with these kind of people. In this problem, with the help of the US Census data, we aim to try to see these kind of traits in the people who earn more than 50,000 US Dollars a year.

The dataset given consists of some characteristics, like the age, working class, marital status, occupation etc., of the working class of the United States of America. From the problem that we are trying to solve, we get an idea of the important factors that lead to a person making more than 50,000 US Dollars in a year. We also try and find some correlations between other features in general. This helps us in better understanding the demography of the working class.

In order to solve this problem, we use the Gaussian Naive Bayes model that is offered by the sklearn package. We use multiple plotting libraries for the visualisations which have been shown through this paper. In order to tackle this problem, we compare two approaches. The first approach is one which tries to fit a model to the whole dataset in one go. The second one is an approach where we split the dataset into two parts (based on some criteria) and fit a model to each of those splits. The results for the same have been depicted in a later section.

Through this study we hope to better understand the demography of the working class and how we can best represent it using the Naive Bayes model. Section II talks about the dataset which we have used for the study along with some visuals that support some insights from it. Section III dives into the working of the Naive Bayes model and the evaluation metrics that we will be using for this problem. Section IV dives into the implementation details, where we talk about both the approaches and we try to visualise the fit of the model and reason out in which scenarios each approach works best.

II. Datasets

The dataset given to us consists of the data of 32,561 people belonging to the working class of the United States of America. The details given to us are - Their Age, Working Class, Final Weight, Level of Education, Number of Years of Education, Marital Status, Occupation, Relationship, Race, Gender, Capital Gains, Capital Losses, Working Hours per Week, Native Country and Whether or not they earn more than 50,000 US Dollars in a year.

We notice that there's no visible null values on a simple inspection. But when we try seeing the unique values of each column, we see that some columns have a "?" symbol present in an entry. This represents a missing value and we replace all such values with the word "Missing". We then try to see if any of the columns which are continuous in nature have any visible distribution.

We find that the Age and Final Weight features approximately follow a Gaussian Distribution as shown in Fig 1. Another interesting thing to note is that when we see a histogram of the number of Hours Worked per Week (as shown in Fig. 2), we notice a sharp spike at the number 40. This is because this is the work hour commitment for a normal day job for any company. This is further validated by the fact that most of the people work in the Private Sector and are not self employed or in a kind of Working Class which allows flexible working hours (as shown in Fig. 3). As can be seen in Fig. 4, the majority of the population are centred around a few key occupations like Prof-Speciality, Craft special, Exec-Managerial, Adm-Clerical and Sales. Capital Gain and Capital Loss are two of the features which contain a large number of zeros. On further inspection, we see that we can divide the population into two kinds of people - Investors (Those who have either a non-zero Capital Gain or Capital Loss) and Non-Investors (Those who have Zero Capital Gain and Zero Capital Loss). We make a new feature called Capital Profit and Loss which is calculated as the difference of the Capital Gain and Capital Loss. We split the dataset into two parts based on whether or not the value of the Capital Profit and Loss is zero. Fig 5, 6 and 7 show the correlation heatmap obtained for the whole population, the Investors and the Non-Investors.

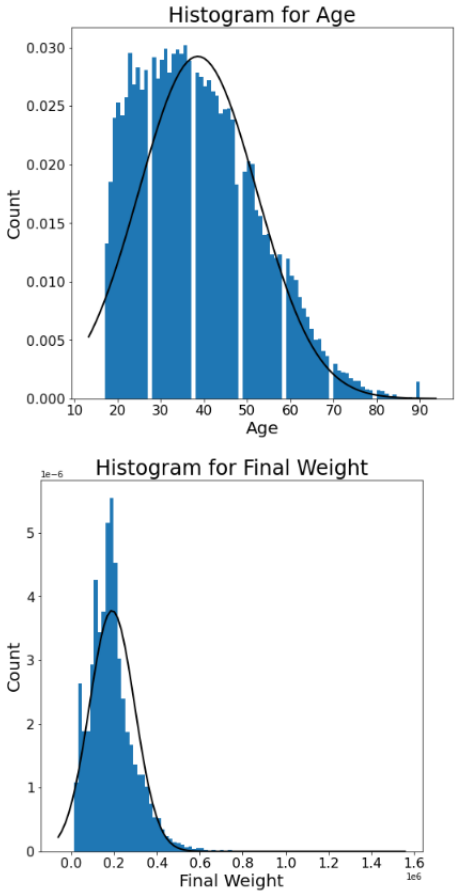

Fig. 1. Gaussian Distributions for Age and Final Weight

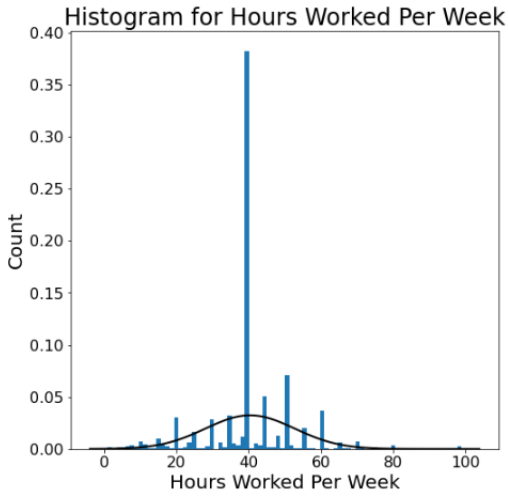

Fig. 2. Histogram for the number of Hours Worked Per Week

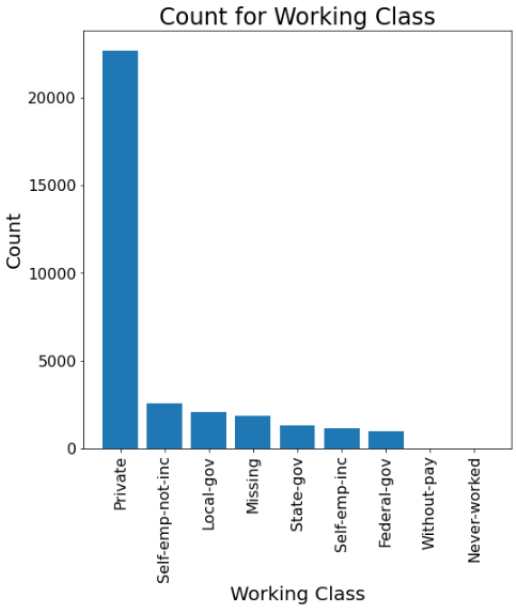

Fig. 3. Working Class Distribution

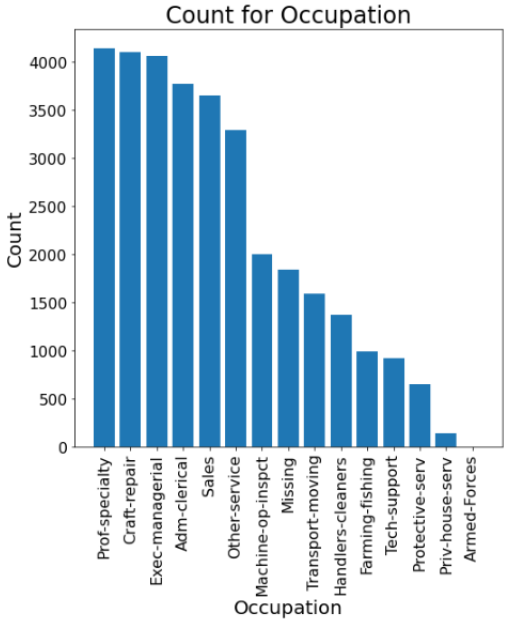

Fig. 4. Occupations taken on by the people

As can be seen in the above graphs, the points that were discussed above are verified. The Age and Final Weight follow a Normal Distribution, there's a huge surge in the Work hour distribution at the 40 Hour mark, this is because of the High Employment in the Private Sector and the majority of the occupations are centred around some key occupations as shown above.

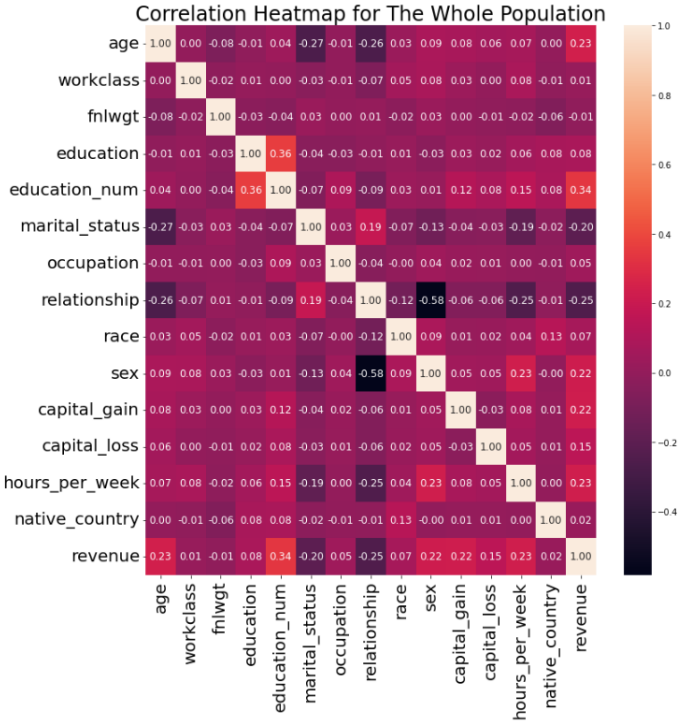

Fig. 5. Population Correlation

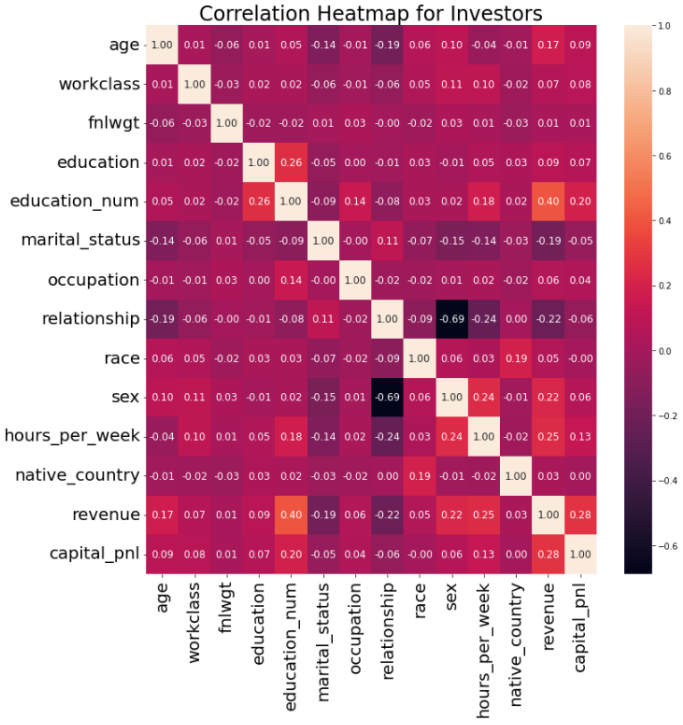

Fig. 6. Investor Correlation

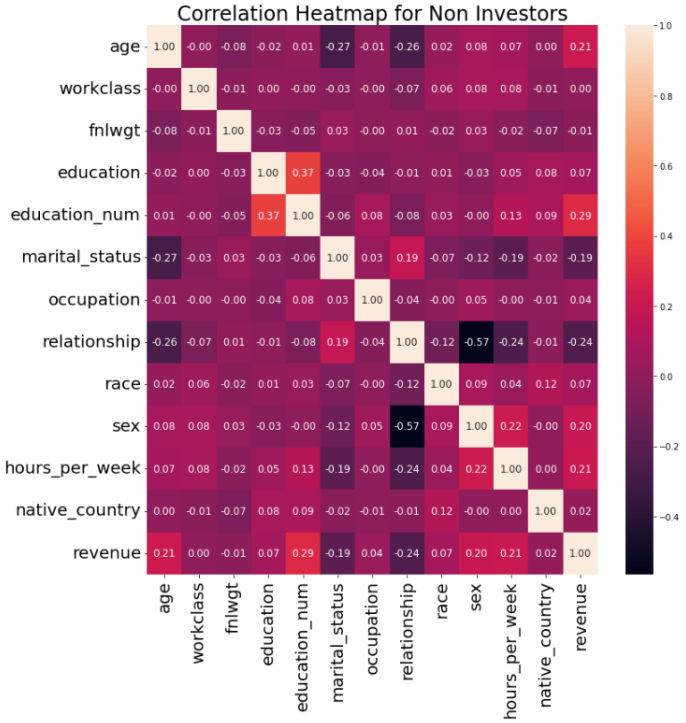

Fig. 7. Non-Investor Correlation

In all the three Figures, we can see that there are some common trends. The Age, Number of Years of Education, Marital Status, Relationship, Gender and Number of Work Hours per Week show a large correlation to the Target Variable which is the Revenue earned.

III. Models

This section takes a deep dive into the mathematical and intuitive concepts involved in the Naive Bayes Algorithm.

A. Intuition

Naive Bayes Classifier uses the Bayes' theorem to predict membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class.

B. The Algorithm

Bayes Theorem states that -

$$P(x|y) = \frac{P(x|y)P(y)}{P(x)} \tag{1}$$

Where,

- P($c|x$): posterior probability of class(c,target) given predictor(x,attributes). This represents the probability of c being true, provided x is true.
- P(c): is the prior probability of class. This is the observed probability of class out of all the observations.

- P($x|c$): is the likelihood which is the probability of predictor-given class. This represents the probability of x being true, provided c is true.
- P(x): is the prior probability of predictor. This is the observed probability of predictor out of all the observations.

Naive Bayes Classifier assumes that all the features are unrelated to each other. Presence or absence of a feature does not influence the presence or absence of any other feature. In real world datasets, we test a hypothesis given multiple evidence on features. So, the calculations become quite complicated. To simplify the work, the feature independence approach is used to uncouple multiple evidence and treat each as an independent one.

C. Types of Naive Bayes Algorithms

There are three major types of Naive Bayes Algorithms. Many modifications of these exist, but the crux of the models remain the same. They are:

- Gaussian Naive Bayes: Which assumes that the underlying distribution of the dataset is Gaussian in nature
- Multinomial Naive Bayes: Which cares about counts for multiple features that do occur.
- Bernoulli Naive Bayes: Which cares about counts for a single feature that do occur and counts for the same feature that do not occur

D. Applications of Naive Bayes

Naive Bayes is one of the most used and fastest classification algorithms. It is very well suited for large volume of data. It is successfully used in various applications such as:

- Spam filtering
- Text Classification
- Sentiment Analysis
- Recommendation Systems

It uses the Bayes theorem of Probability for Prediction of Unknown Class.

IV. Modelling

In this section, we contrast and compare two of our approaches. One where we fit a model to the whole population and one where we have two models - One for the investors and one for the Non-Investors. The vanilla model achieves an F1-Score of 0.33. When we split the decision making criteria, we see that the Investor model achieves an F1-score of 0.74 and the Non-Investor Model that of 0.34, both of which are improvements over the vanilla model. This shows that making this split in predictions will help make better decisions regarding the demography. Given in Fig 8 - 12 are the parameter values learned by the Naive Bayes model for the various cases. We fit a Gaussian Naive Bayes for the Vanilla Model and the Investor Model and we use a Multinomial Naive Bayes for the Non-Investors.

	Revenue <= 50K	Revenue >50K
age	36.773244	44.162882
workclass	3.878490	3.908858
fnlwgt	190242.214153	187374.587908
education	10.133019	10.853361
education_num	9.583916	11.616784
marital_status	2.788034	2.078057
occupation	6.330432	6.791397
relationship	1.678604	0.721161
race	3.627859	3.771545
sex	0.611452	0.852609
hours_per_week	38.850002	45.523387
native_country	37.037809	37.460972
capital_pnl	96.340229	3843.405775

Fig. 8. Mean for the Features of the Vanilla Model

	Revenue <= 50K	Revenue >50K
age	2.069172e+02	1.221455e+02
workclass	1.290551e+01	1.351802e+01
fnlwgt	1.126458e+10	1.049606e+10
education	2.816591e+01	1.902911e+01
education_num	1.700481e+01	1.676584e+01
marital_status	1.372054e+01	1.181185e+01
occupation	2.974118e+01	3.037753e+01
relationship	1.347789e+01	1.354272e+01
race	1.186135e+01	1.162498e+01
sex	1.131902e+01	1.120711e+01
hours_per_week	1.622374e+02	1.304965e+02
native_country	5.403683e+01	4.663598e+01
capital_pnl	1.025130e+06	2.157017e+08

Fig. 9. Standard Deviation for the Features of the Vanilla Model

	Revenue <= 50K	Revenue >50K
age	40.707702	45.082812
workclass	3.818960	4.004333
fnlwgt	186038.717577	188183.861338
education	10.271231	10.838710
education_num	9.780777	11.946558
marital_status	2.634628	2.110737
occupation	6.356814	6.868079
relationship	1.468729	0.735195
race	3.691244	3.776119
sex	0.647136	0.848339
hours_per_week	39.830151	45.860857
native_country	37.030283	37.461242
capital_pnl	1336.330481	12227.597978

Fig. 10. Mean for the Features of the Investor Model

	Revenue <= 50K	Revenue >50K
age	2.395507e+02	1.295570e+02
workclass	1.237823e+01	1.242134e+01
fnlwgt	1.035463e+10	1.002059e+10
education	2.483711e+01	1.813632e+01
education_num	1.659863e+01	1.598519e+01
marital_status	1.286211e+01	1.102297e+01
occupation	2.992622e+01	2.865191e+01
relationship	1.262382e+01	1.256549e+01
race	1.083049e+01	1.069575e+01
sex	1.039117e+01	1.029148e+01
hours_per_week	1.728410e+02	1.343625e+02
native_country	5.801970e+01	4.485234e+01
capital_pnl	1.359433e+07	5.841815e+08

Fig. 11. Standard Deviation for the Features of the Investor Model

	Revenue <= 50K	Revenue >50K
age	-8.563115	-8.361959
workclass	-10.802544	-10.783542
fnlwgt	-0.000791	-0.000889
education	-9.847285	-9.759576
education_num	-9.900871	-9.706207
marital_status	-11.134138	-11.418490
occupation	-10.306366	-10.231865
relationship	-11.637592	-12.455309
race	-10.871057	-10.813991
sex	-12.652151	-12.306726
hours_per_week	-8.502400	-8.329458
native_country	-8.548577	-8.525213

Fig. 12. Log Probabilities for the Features of the Non Investor Model

V. Conclusions

In this project, I build a Gaussian Naive Bayes Classifier model to predict whether a person makes over 50K a year. We compare two approaches, one where we fit a model to the whole dataset, called the Vanilla Approach and one where we split the dataset into two parts - One consisting of people who invest their money and one consisting of people who do not invest their money. As seen, the second approach significantly outperforms the first one in terms of our performance metrics. We conclude that this approach is a better one for decision making overall, since the characteristics of the people belonging to these two groups are significantly different. If a single model is used to fit moth, it will miss out on the unique behaviours of these two groups which is exactly what we observe in our experiments.

VI. Avenues For Further Research

One possible area of improvement would be to have a different model for the categorical features (a Multinomial Naive Bayes) and another for the Continuous ones (a Gaussian Naive Bayes). Combining the Predicted probabilities of these models for our prediction, should increase the reliability of our prediction framework.

References

[1] An Article on Gaussian Naive Bayes: https://towardsdatascience.com/gaussian-naive-bayes-4d2895d139a

[2] An Article on Multinomial Naive Bayes: https://www.upgrad.com/blog/multinomial-naive-bayes-explained/