A Mathematical Essay on Naive Bayes Classifier

Jeshlin Donna J
Department of Metallurgical and Materials Engineering
Indian Institute of Technology Madras
Chennai, India
mm20b029@smail.iitm.ac.in

***Abstract*—The data considered was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The key task is to determine whether a person makes over $50K a year. The aim is use and apply Naive Bayes Classifier and perform Exploratory data analysis to uncover previously unknown or hidden facts in the data set available, visualize them using Data Visualization and make predictions on the validation data. This is followed by model evaluation on the Naive Bayes Classifier model. The task is to conclude the study of which types of individuals are more likely to make over $50K a year.**

***Index Terms*—Naive Bayes Classifier, Exploratory data analysis, Data Visualization, Model Evaluation**

I. INTRODUCTION

The data considered was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The key task is to determine whether a person makes over $50K a year by using the supervised learning technique of Naive Bayes Classifier.

A Naive Bayes classifier is a probabilistic machine learning model that is used for classification tasks. It is a classification technique based on Bayes Theorem with an assumption of independence among predictors i.e. a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Naive Bayes classifiers are known to outperform and can be extremely fast when compared to the more sophisticated classification methods.

The goal of this research paper is to correctly predict who are more likely to make over $50K a year given a set of information. A predictive classification model based on Naive Bayes classifier was built using the given data so people's 'Work Class', 'Education', 'Marital status', 'Occupation', 'Race', 'Sex', 'Native Region', 'Capital gain', 'Capital loss' and many other features contributed to whether they would make over $50K a year. The output of whether a person will make over $50K a year is predicted using the Naive Bayes classifier based on different feature combinations of the data.

In this paper, we conducted exploratory data analysis to excavate different knowledge existing in the given data set and to perceive the impact of every field with respect to the chances of making over $50K a year by the use of ">50k" field analysis in between each field of the data set. Data analysis was performed, feature engineering was applied on the data and, a Naive Bayes Classifier model was built and trained on the training data, and the accuracy was tested on the validation dataset. After analyzing the dataset, we discovered insights and information on what the people who made over $50K a year had in common that helped them do so. We were also able to predict if a particular person with certain feature values would make over $50K a year by applying the tools of machine learning and statistical analysis.

II. NAIVE BAYES CLASSIFIER

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifiers have high accuracy and speed on large datasets.

Abstractly, Naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $x = (x_1, \ldots, x_n)$ representing some n features (independent variables), it assigns to this instance probabilities: $p(C_k \mid x_1, \ldots, x_n)$ for each of K possible outcomes or "classes" C_k.

The problem with the above formulation is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on conditional probability is infeasible. The model must therefore be reformulated to make it more tractable. Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x} \mid C_k)}{p(\mathbf{x})} \tag{1}$$

In plain English, using Bayesian probability terminology, the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \tag{2}$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on C and the values of the features x_i are given, so that the denominator is effectively constant.

The numerator is equivalent to the joint probability model $p(C_k, x_1, \ldots, x_n)$, which can be rewritten as $p(x_1 \mid x_2, \ldots, x_n, C_k)\ p(x_2 \mid x_3, \ldots, x_n, C_k) \cdots p(x_{n-1} \mid x_n, C_k)\ p(x_n \mid C_k)\ p(C_k)$ using the Chain rule for repeated applications of the definition of conditional probability.

Now the "naive" conditional independence assumptions come into play: assume that all features in x are mutually independent, conditional on the category C_k. Under this assumption, $p(x_i \mid x_{i+1}, \ldots, x_n, C_k) = p(x_i \mid C_k)$.

Thus, the joint model can be expressed as

$$p(C_k \mid x_1, \ldots, x_n) \propto p(C_k, x_1, \ldots, x_n) \tag{3}$$

$$\propto p(C_k)\ p(x_1 \mid C_k)\ p(x_2 \mid C_k)\ p(x_3 \mid C_k) \cdots \tag{4}$$

$$\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k), \tag{5}$$

This means that under the above independence assumptions, the conditional distribution over the class variable C is:
$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$,
where the evidence $Z = p(\mathbf{x}) = \sum_k p(C_k)\ p(\mathbf{x} \mid C_k)$ is a scaling factor dependent only on $x_1, \ldots, x_n$, that is, a constant if the values of the feature variables are known.

A. Constructing a classifier from the probability model

The discussion so far has derived the independent feature model, that is, the naive Bayes probability model. The naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the 'maximum a posteriori' or 'MAP' decision rule. The corresponding classifier, a Bayes classifier, is the function that assigns a class label $\hat{y} = C_k$ for some k as follows:

$$\hat{y} = \underset{k \in \{1, \ldots, K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

B. Parameter estimation and event models

A class's prior may be calculated by assuming equiprobable classes (i.e., $p(C_k) = 1/K$), or by calculating an estimate for the class probability from the training set (i.e., prior for a given class = number of samples in the class/total number of samples). To estimate the parameters for a feature's distribution, one must assume a distribution or generate nonparametric models for the features from the training set.

The assumptions on distributions of features are called the "event model" of the Naive Bayes classifier. For discrete features like the ones encountered in document classification (include spam filtering), Multinomial and Bernoulli distributions are popular. These assumptions lead to two distinct models, which are often confused.

1) Gaussian naïve Bayes: When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. For example, suppose the training data contains a continuous attribute, 'x'. The data is first segmented by the class, and then the mean and variance of x is computed in each class. Let μ_k be the mean of the values in x associated with class 'C_k', and let σ_k^2 be the Bessel corrected variance of the values in x associated with class 'C_k'. Suppose one has collected some observation value v. Then, the probability 'density' of v given a class C_k, $p(x = v \mid C_k)$, can be computed by plugging v into the equation for a normal distribution parameterized by μ_k and σ_k^2. That is,

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \tag{6}$$

Another common technique for handling continuous values is to use binning to discretize the feature values, to obtain a new set of Bernoulli-distributed features; some literature in fact suggests that this is necessary to apply naive Bayes, but it is not, and the discretization may throw away discriminative information.

Sometimes the distribution of class-conditional marginal densities is far from normal. In these cases, kernel density estimation can be used for a more realistic estimate of the marginal densities of each class. This method, can boost the accuracy of the classifier considerably.

2) Multinomial naïve Bayes: With a multinomial event model, samples (feature vectors) represent the frequencies with which certain events have been generated by a Multinomial distribution $(p_1, \ldots, p_n)$ where p_i is the probability that event i occurs (or K such multinomials in the multiclass case). A feature vector $\mathbf{x} = (x_1, \ldots, x_n)$ is then a histogram, with x_i counting the number of times event i was observed in a particular instance. The likelihood of observing a histogram 'x' is given by

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} {p_{ki}}^{x_i} \tag{7}$$

The multinomial naïve Bayes classifier becomes a linear classifier when expressed in log-space:

$$\log p(C_k \mid \mathbf{x}) \propto \log \left(p(C_k) \prod_{i=1}^{n} {p_{ki}}^{x_i} \right) \tag{8}$$

$$= \log p(C_k) + \sum_{i=1}^{n} x_i \cdot \log p_{ki} \tag{9}$$

$$= b + \mathbf{w}_k^{\top} \mathbf{x} \tag{10}$$

where $b = \log p(C_k)$ and $w_{ki} = \log p_{ki}$.

If a given class and feature value never occur together in the training data, then the frequency-based probability estimate will be zero, because the probability estimate is directly proportional to the number of occurrences of a feature's value. This is problematic because it will wipe out all information in the other probabilities when they are multiplied. Therefore, it is often desirable to incorporate a small-sample correction, called pseudocount, in all probability estimates such that no probability is ever set to be exactly zero. This way of regularizing naive Bayes is called Lidstone smoothing.

3) Bernoulli naïve Bayes: In the multivariate Bernoulli event model, features are independent Booleans (binary variables) describing inputs.

If x_i is a boolean expressing the occurrence or absence of the i'th term from the vocabulary, then the likelihood of a document given a class C_k is given by $p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i}(1 - p_{ki})^{(1-x_i)}$, where p_{ki} is the probability of class C_k generating the term x_i.

III. USING NAIVE BAYES CLASSIFIER TO SOLVE THE PROBLEM

A. Description of the Dataset

The data considered was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics).

The dataset comprises of 32560 observations and 15 variables. Out of which one variable('<=50K') is the dependent variable and the rest 14 are independent variables.

```
RangeIndex: 32560 entries, 0 to 32559
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   39              32560 non-null  int64
 1    State-gov      32560 non-null  object
 2   77516           32560 non-null  int64
 3    Bachelors      32560 non-null  object
 4   13              32560 non-null  int64
 5    Never-married  32560 non-null  object
 6    Adm-clerical   32560 non-null  object
 7    Not-in-family  32560 non-null  object
 8    White          32560 non-null  object
 9    Male           32560 non-null  object
 10  2174            32560 non-null  int64
 11  0               32560 non-null  int64
 12  40              32560 non-null  int64
 13   United-States  32560 non-null  object
 14   <=50K          32560 non-null  object
dtypes: int64(6), object(9)
```

Fig. 1. Information about the dataset

It was noted that there is notably a large difference between 75th %tile and maximum values (Fig.2.) in most of the variables. This suggests that there are extreme values(outliers) in the data.

	39	77516	13	2174	0	40
count	32560.000000	3.256000e+04	32560.000000	32560.000000	32560.000000	32560.000000
mean	38.581634	1.897818e+05	10.080590	1077.615172	87.306511	40.437469
std	13.640642	1.055498e+05	2.572709	7385.402999	402.966116	12.347618
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.178315e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783630e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.370545e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

Fig. 2. Statistics of the dataset

B. Preparing and cleaning the data

The raw dataset was found to be having vague column/variable names, these variables were renamed with meaningful variable names after performing a background research on the dataset. The categorical variable ' <=50K' was mapped to numerical binary values (i.e. 0/1) to create a new binary variable '>50K'. Few rows were found to be containing meaningless/missing data, these rows were removed from the dataframe.

C. Feature Engineering

New variables called "Working hour level", "capital_gain", "capital_loss" and "Native Region" have been created.

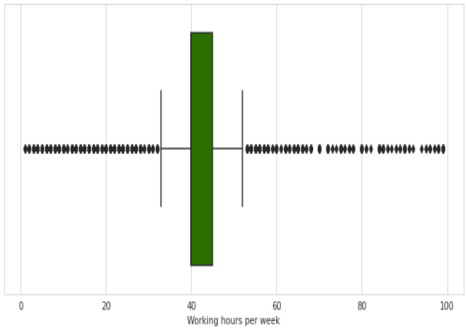

Fig. 3. Box plot of the variable "Working hours per week"

1) Creating a new variable called "Working hour level": From the the box plot of the variable "Working hours per week"(Fig.3.), we see that the mean number of working hours per week is around 40 and at least 50% of the people taking part in the survey, work between 40 and 45 hours per week.

Therefore, the working hours was grouped into 5 categories which was considered relevant: less than 40 hours per week, between 40 and 45 hours per week, between 45 and 60 hours per week, between 60 and 80 hours per week, and more than 80 hours per week. A new a new categorical variable called "Working hour level" was created with 5 levels corresponding to these 5 categories.

2) Creating a new variable called "Native Region": The factor/categorical variable "Native Country" had 41 different categories. If a predictive model was built with "Native Country" as a covariate, it would have resulted in 41 additional degrees of freedom due to this categorical variable.

This will unnecessarily complicate the analysis and might lead to overfitting. Hence, the native countries were grouped into several global regions. This coarsening of the data also made the interpretation of the results easier to comprehend.

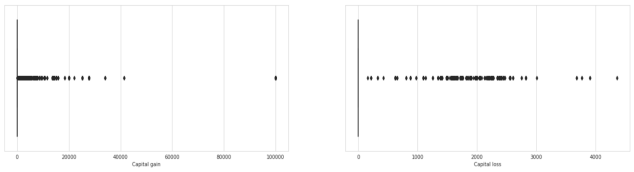

Fig. 4. Boxplot for "Capital gain" and "Capital loss" variables

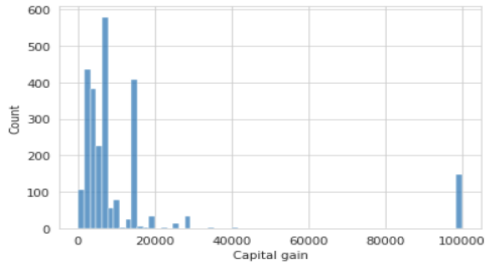

Fig. 5. Histogram for Non-zero Capital gain

3) Creating new variables called "capital_gain" and "capital_loss: The percentage of zero values in the "Capital gain" and "Capital loss" variables were found to be approximately 91.59% and 95.27% respectively. This can seriously disrupt the data analysis.

Based on the box plots(Fig.4.) and histograms(Fig.5.) that illustrate visually the results of the summary statistics for the non-zero capital gain and capital loss, the values of the variables "capital_loss" and "capital gain" have been grouped into categories and two new factor variables called "capital_gain" and "capital_loss" were created.

Grouping was done in the following way:

Capital gain: All values of "capital_gain" which are less than the first quartile of the nonzero capital gain as "Low"; all values that are between the first and third quartile as "Medium"; and all values greater than or equal to the third quartile are marked as "High".

Capital loss: All values of "capital_loss" which are less than the first quartile of the nonzero capital gain as "Low"; all values that are between the first and third quartile - as "Medium"; and all values greater than or equal to the third quartile are marked as "High".

D. Visualizations and insights from the analysis

From (Fig.6.) we infer that, people who are employed under 'Incorporated self employment' have the highest probability of making >50K followed by those who's work class is 'Fedral Government'. People who are employed under the 'Private' work class have the least probability of making >50k.

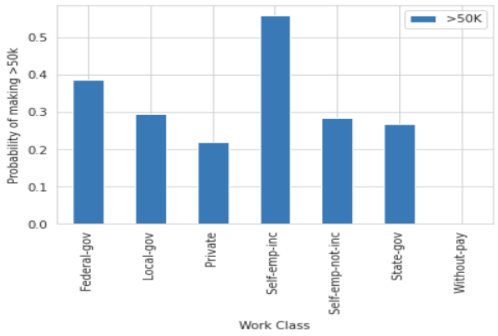

Fig. 6. Probability of making >50k for the different Work Classes

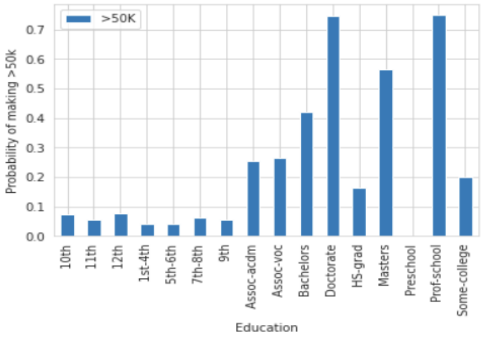

Fig. 7. Probability of making >50k for different education levels

From (Fig.7.) we infer that, people who had completed their education upto a Professor-School, Doctorate or Masters degree(in the decreasing order of mention) had a higher probaibility of earning well. Whereas people who had only completed their education upto the schooling level are expected to make significantly less when compared to the others.

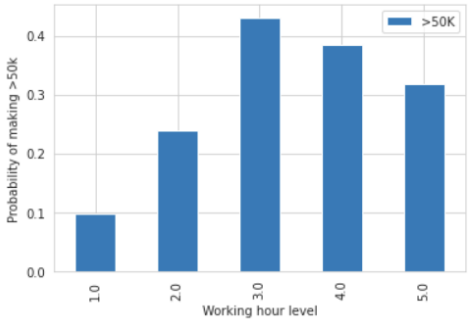

Fig. 8. Probability of making >50k for the various working hour levels

From (Fig.8.) we can conclude that, the people who worked between 45-60(45 not included) hours per week(working hour level 3) had highest probability of making >50k, followed by

those who worked between 60-80(60 not included) hours per week(working hour level 4). Higher working hours does not accurately mean higher payoffs. (i.e. working hour level 5 has lesser probability of paying off well when compared to level 3 and 4).

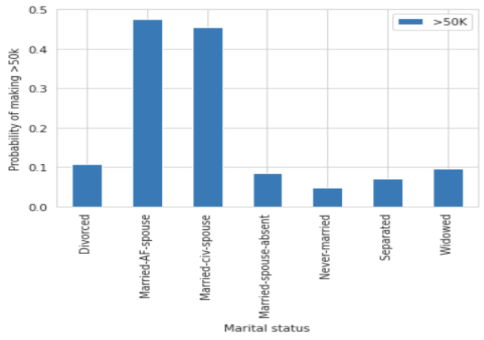

Fig. 9. Probability of making >50k with respect to Marital Status

From(Fig.9.) we can conclude that, people who were married to a spouse in the Armed Forces or a civilian spouse had a higher probability of making >50k. People who were never married had the least probability of the same.

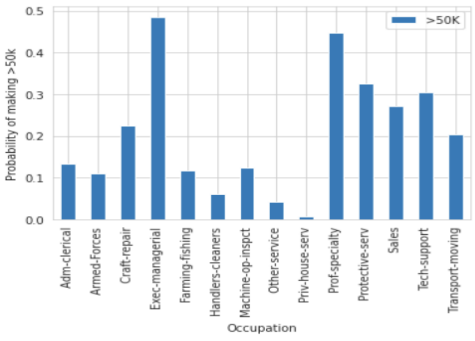

Fig. 10. Probability of making >50k for various occupations

From(Fig.10.) we can infer that the people who were employed as executive managers, professional speciality and protective services made considerably more than the others. People employed under private house services, machine operations, farmimg, fishing and cleaning services made significantly lesser than the rest.

From (Fig.11.), it can be said to be a general inference that people who make >50k tend to have older age than those who make <50k.

From(Fig.12.) it is evident that the probability of men making >50k is found to be much higher than that of female. This might be due to the fact that in the 1900's women were barely allowed to work, and even if they did they were neither given work in the positions of higher order nor paid equally well as their male counterparts.

From (Fig.13.), we see that the people belonging to the race 'Asian Pacific Islander' and 'White' have a higher probability of making >50k when compared to the 'Black', 'Indian-American', 'Eskimo' or other groups.

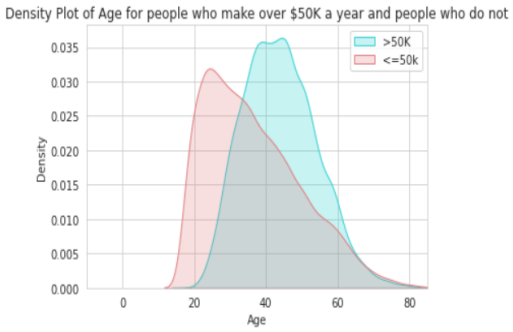

Fig. 11. Density Plot of Age for people who make over $50K a year and people who do not

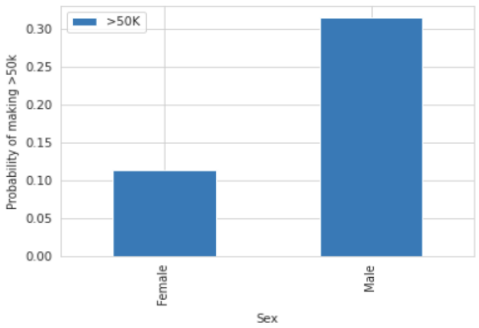

Fig. 12. Probability of making >50k for male and female

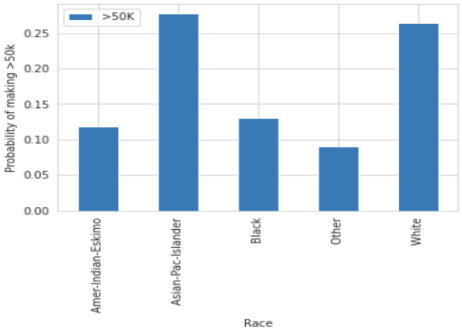

Fig. 13. Probability of making >50k for different races

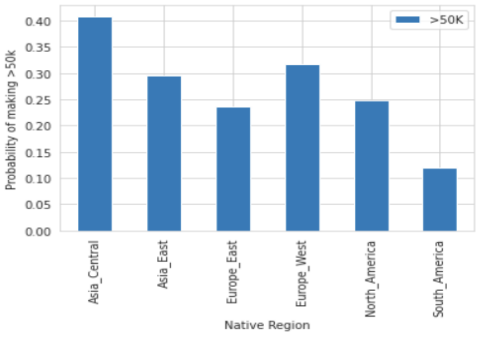

Fig. 14. Probability of making >50k for different Native Regions

From (Fig.14.),we see that the people natively belonging to Central Asia and Western Europe have the highest probability of making >50k. Whereas those who's nativity is from Southern America have the least probability of making >50k.

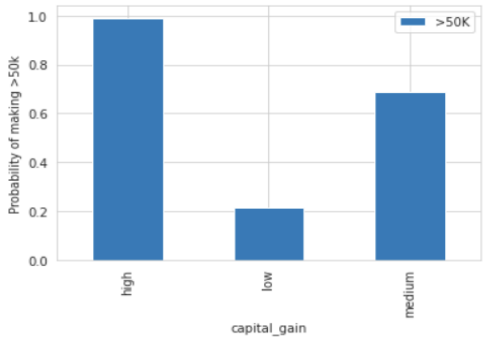

Fig. 15. Probability of making >50k for different levels of Capital Gain

From the histographic plot(Fig.15.), it is evident that those people with a higher capital gain are generally more likely to make >50k.

E. Data Preprocessing

In the preprocessing stage, One-Hot Encoding was performed on the categorical variables 'Work Class', 'Education', 'Marital status', 'Occupation', 'Relationship', 'Race', 'Sex', 'Native Region', 'capital_gain' and 'capital_loss'.

	Age	fnlwgt	Years of education	>50K	Working hour level
count	30161.000000	3.016100e+04	30161.000000	30161.000000	30161.000000
mean	38.437883	1.897976e+05	10.121216	0.248931	2.045589
std	13.134882	1.056527e+05	2.549983	0.432401	0.764435
min	17.000000	1.376900e+04	1.000000	0.000000	1.000000
25%	28.000000	1.176280e+05	9.000000	0.000000	2.000000
50%	37.000000	1.784290e+05	10.000000	0.000000	2.000000
75%	47.000000	2.376300e+05	13.000000	0.000000	2.000000
max	90.000000	1.484705e+06	16.000000	1.000000	5.000000

Fig. 16. Statistics of the numerical non-categorical features in the data

It was observed that the numeric columns in our dataset have varying ranges(Fig.16.). Thus, the numerical variables- 'Age', 'fnlwgt', 'Years of education', '>50K' and 'Working hour level' were scaled using the Min-Max Scaler from the Ski-kit Learn library. Such scaling of the numeric features ensures that no particular feature has a disproportionate impact on the model's loss.

The correlation between the variables "Working hours per week" and "Working hour level" was found to be very

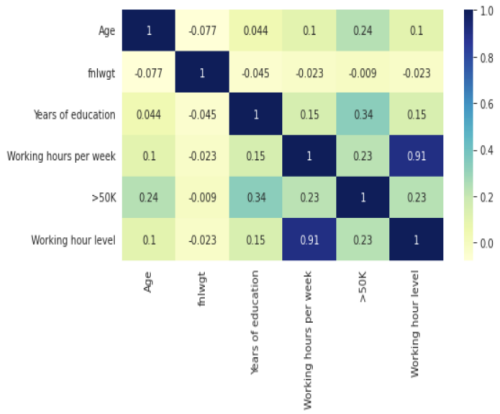

Fig. 17. Correlation heat-map between the various features of data

high(Fig.17.). Hence the former variable was dropped before training to remove redundancy.

The dataset was also split into the training and validation parts with the validation data being used to calculate the accuracy and performance of the model before making predictions on the test data.

F. Building the Naive Bayes Classifier

A Naive Bayes Classifier has been built using the inbuilt method from the Skikit-Learn library. The dataset was split into the training and validation parts and the training dataset was used to train the model to fit the given data. The model was found to fit the training data with a 67.04% accuracy.

G. Model performance on the Validation dataset

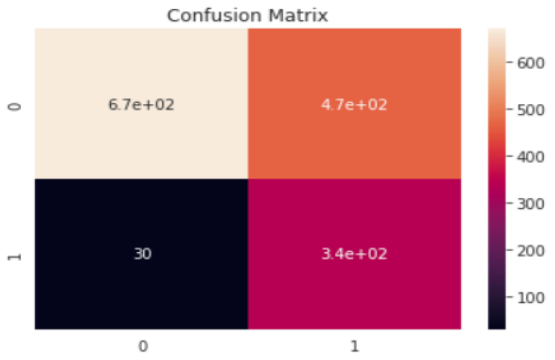

Fig. 18. Confusion matrix of our model on the validation dataset

The performance of the Naive Bayes Classifier was evaluated on the validation dataset. The accuracy of the model was found to be 67.0%. And a confusion matrix of our model(Fig.18.) on the validation dataset was used to visualize the miss-classified datapoints.

IV. CONCLUSIONS

We analyzed the role of adult census data in affecting the chances of making >50k. People who were employed under 'Incorporated self employment' were found to have the highest probability of making >50K followed by those who's work class is 'Fedral Government'. People who are employed under the 'Private' work class have the least probability of making >50k. People who had completed their education upto a Professor-School, Doctorate or Masters degree(in the decreasing order of mention) had a higher probability of earning well. Whereas people who had only completed their education upto the schooling level are expected to make significantly less when compared to the others. People who worked between 45-60(45 not included) hours per week had highest probability of making >50k, followed by those who worked between 60-80(60 not included) hours per week. But it was also inferred that higher working hours did not necessarily mean higher payoffs. Those people with a higher capital gain were found to be more likely to make >50k. It was also concluded that people who were married to a spouse in the Armed Forces or a civilian spouse had a higher probability of making >50k. People who were never married had the least probability of the same. With regards to occupation, the people who were employed as executive managers, in professional speciality and protective services made considerably more than the others. People employed under private house services, machine operations, farming, fishing and cleaning made significantly lesser than the rest. It has also been evidently discovered that the probability of men making >50k is found to be much higher than that of female. This might be due to the fact that in the 1900's women were barely allowed to work, and even if they did they were neither given work in the positions of higher order nor paid equally well as their male counterparts. We see that the people belonging to the race 'Asian Pacific Islander' and 'White' have a higher probability of making >50k when compared to the 'Black', 'Indian-American', 'Eskimo' or other groups. We also see that the people natively belonging to Central Asia and Western Europe have the highest probability of making >50k. Whereas those who's nativity is from Southern America have the least probability of making >50k.

Further improvements in the model are highly possible. In the present model, outliers found in the data were not accounted for, further work can done on addressing the outliers present in the data. More deeper techniques of feature engineering can also be used to extract the best possible inferences from the data variables. Feature grouping can be tried out for more variables to further reduce the chances of over-fitting and to prevent the model from learning the noise in the data. Better techniques can be tried out to handle the significant amount of zero values in the capital loss and capital gain variables. In order to further improve the overall result, an extensive hyper-parameter tuning should also be done to improve the accuracy of the model for it to better fit the data. It is possible to further improve it by doing ensemble learning or finding better machine learning algorithms that fit the data better.

REFERENCES

[1] Rish, Irina. (2001). An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods Artif Intell. 3.

[2] Berrar, Daniel. (2018). Bayes' Theorem and Naive Bayes Classifier. 10.1016/B978-0-12-809633-8.20473-1.

[3] Nayak, Nikhil. (2020). Application of Naive Bayes Classifier for Information Extraction. 10.31219/osf.io/z7q2e.