Predicting the Eligibility of H-1B Visa applications using neural networks

Apurva Katti
Spring 2019
CIS 5930 Advanced Data Mining
Department of Computer Science
Florida State University

***Abstract*—H-1B Visa petitions are filed every year by foreign nationals to work in the United States. But there exists a limit on the number of visas issued each year.This project considers the prediction of the eligibility of H-1B Visa applications(petitions) depending on the various parameters on the applicants list.The main aim is to learn this as a classification problem by extracting the important features of the data.The input to the model are the features of the applicant and since it is a supervised learning problem, the output is a class label making the prediction.Data is pre-processed to the requirement and a Multilayer Perceptron(MLP) model is trained with dense and dropout layers.A through analysis of the model is performed and results are plotted.The results indicate that the model has the capacity to attain upto 94% accuracy when trained as a multiclass classification problem.**

I. Introduction

H-1B is a type of non-immigrant visa in the United States that allows foreign nationals to work in occupations that require specialized knowledge and a bachelor's degree or higher in the specific specialty.The duration of the stay is three years, extendable up to six years.Once the Visa expires the applicant is expected to renew by re-applying it. There is also a limit on the number of H-1B visas issued each year and a total of 180,440 new and initial H1-B visas were issued in 2017. The Visa expects that the applicant has a job offer from a legitimate employer before applying for the Visa.

However the US immigration service (USCIS) allocates the Visa to applicants based on a lottery system and therefore the characteristics which determine the process of selection is not known.In such a scenario, given the parameters of an applicant, predicting the likelihood of eligibilty of the Visa can become useful.

There is a procedure for filing the visa petitions.The first step of the H1B application process is for the U.S. employer to file the H1B petition on behalf of the foreign employee.The second step is that the prevailing and actual wages should be confirmed by the State Employment Security Agency.The wages of the employee is one of the most significant feature of the applicant which determines where he/she is eligible to obtain the visa.If the prevailing wage exceeds the offer made by the employer then necessary steps will be made to calculate the actual wage. The third step of the H-1B application process is to file the Labor Condition Application. The next step is to prepare the petition and file it at the proper USCIS office. With the approved labor application, the employer then files a I-129 form requesting H-1B classification for the employee. The final step of the H1B application process is to check the status of your H1B visa petition.

Each year the USCIS publishes a memo when enough subject applications have been received, indicating the closure of subject application season during the month of April and start randomly selecting the applicants based off of the lottery scheme.The applicants with a Master's degree have a higher chance as first a lottery is held to award the 20,000 visas available to master's degree holders, and those not selected are then entered in the regular lottery for the other 65,000 visas.

This work deals with making thorough analysis of the intermediate step before the application is picked up and granted the visa.The intermediate process checks the eligibility of the applicant depending on the various parameters and then issues certified or denied as its status.It also has several other types of statuses such as certified-withdrawn and withdrawn.This is an important problem as there is no clear rules or criteria for rejecting or accepting a petition.

II. Literature Survey

This area of research is not well studied by researches and hence this work might be one of the first to train a neural network on the H-1B petition data.There are some works who have performed several classifiers on the data.Usually they have approached this problem as a classification problem and supervised learning where they are parameters or features which directly affect the prediction label or parameter.Because it is a classification problem one can think of applying different types of classifiers like XGBoost,Random forest and so on or the data can also be studied to make comprehensive analysis such as the number of petitions accepted in each state, top 10 jobs, top sponsoring employers or even the best job.

This problem could also be studied as an unsupervised problem and then perform different clustering algorithms.

III. Background

Since the work employs the concepts of machine learning, it is important to learn them.Neural Networks are complex multi-layer structures that are composed of smaller working units called "neurons". Each neuron introduces a non-linearity with a chosen activation function g(x) along with the initial parameters such as weights and propagates the same to the next layer. At the output layer, a prediction parameter $\hat{y}$ makes the prediction of the problem considered. The cost function is used to minimize the training loss.The parameters, w and b are updated as in using back-propagation.

A. *MultiLayer Perceptron Model:*

As the name suggests, it is a cascading of multiple single layer perceptrons(neurons).It has an input layer that connects to the input variables, one or more hidden layers, and an output layer that produces the output variables.

- **Input Layer** - Input variables, sometimes called the visible layer.
- **Hidden Layer** - Layers of nodes between the input and output layers. There may be one or more of these layers.
- **Output Layer** - A layer of nodes that produce the output variables.

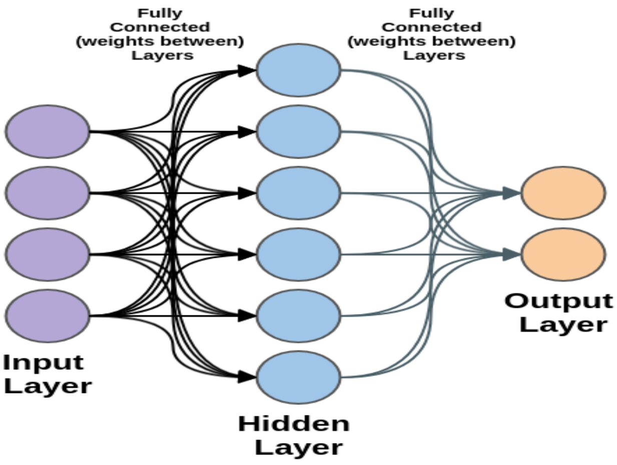

Fully Connected (weights between) Layers
Fully Connected (weights between) Layers
Input Layer
Hidden Layer
Output Layer

Figure 1: MultiLayer Perceptron Model [1]

B. *Back Propagation*

The Backpropagation algorithm is a supervised learning method for multilayer feed-forward networks.The principle of Backpropagation is to model a given function by modifying internal weights of input signals to produce an expected output signal. The system is trained using a supervised learning method, where the error between the system's output and a known expected output is found.Mainly Backpropagation is used to calculate the weights and biases.Backpropagation can be used for both classification and regression problems.

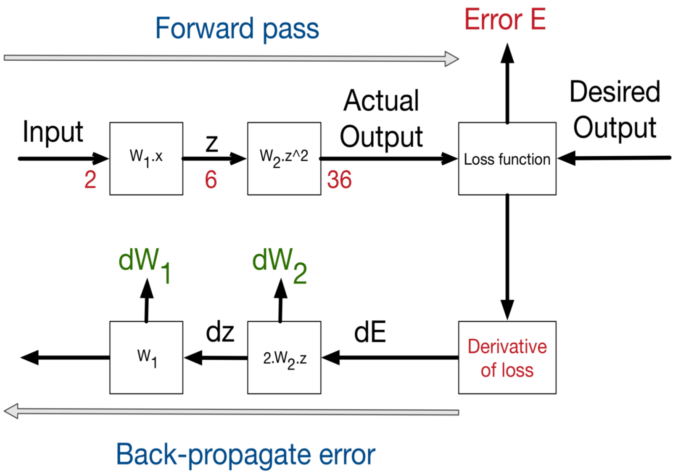


Figure 2: Backpropagation [1]

C. *Activation function*

The activation functions are responsible for transforming the summed weighted input from the node into the activation of the node or output for that input.There are many useful functions such as tanh, relu, sigmoid etc.For example,rectified linear unit returns 0 if it receives any negative input, but for any positive value x it returns that value back. So it can be written as f(x)=max(0,x).The graph can be represented as below.Sftmax function is a type of activation function which provides the probability that any class is true.The values together add up to 1.Since the project deals with multi classes, softmax is used.Examples of activations fucntion graphs with their formula is given in the following pictures.

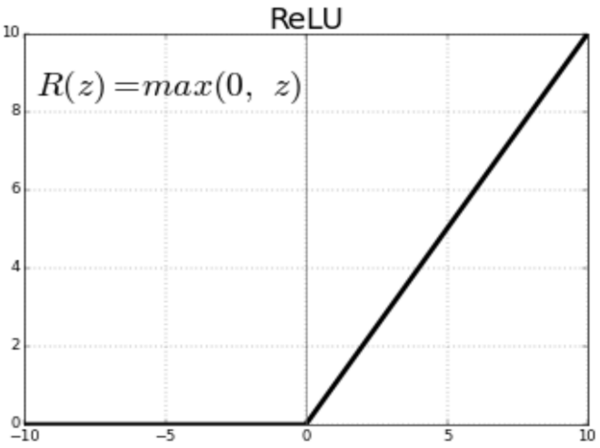


Figure 3: ReLU(Rectified Linear Unit) [3]

D. *Loss function*

Loss functions are used to find the candidate solution to maximize or minimize an optimization algorithm.Usually with neural networks, it is used to minimize the error factor.For a classification problem which involves mapping input variables to a output label, the loss function can be modeled as predicting the probability of an example belonging to each class. Therefore, under maximum likelihood estimation, a set of model weights that minimize the difference between the model's predicted probability distribution given the dataset and the distribution of probabilities in the training dataset. This is called the cross-entropy.

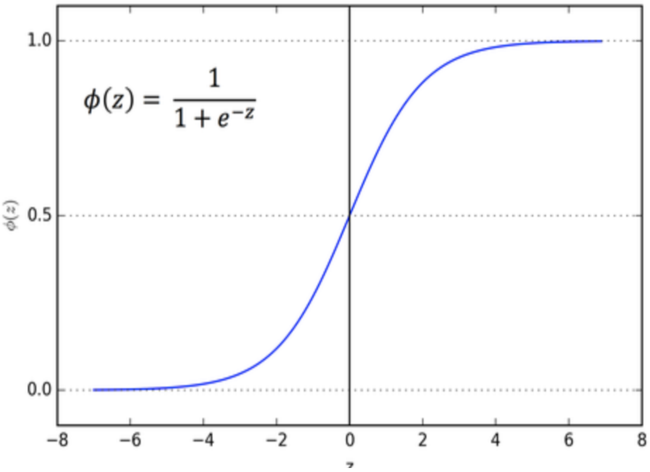


Figure 4: Sigmoid [3]

*1) **Categorical Cross Entropy**:* It is a Softmax activation plus a Cross-Entropy loss.

entropy loss.png

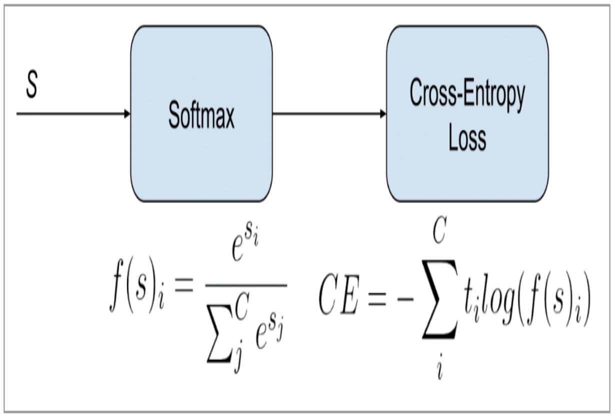

S
Softmax
Cross-Entropy Loss
$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}}$
$CE = -\sum_i^C t_i log(f(s)_i)$

Figure 5: Cross Entropy

*2) **Mean Squared Error**:* Mean Squared Error loss, or MSE is calculated as the average of the squared differences between the predicted and actual values.The result is always positive regardless of the sign of the predicted and actual values and a perfect value is 0.0.

E. One-Hot Encoding

A one hot encoding is a representation of categorical variables(textual form) as vectors(numbers in the range of (0,1)).Consider an example labels with the values 'red' and 'green'.Next, we can create a binary vector to represent each integer value. The vector will have a length of 2 for the 2 possible integer values.The 'red' label encoded as a 0 will be represented with a binary vector [1, 0] where the zeroth index is marked with a value of 1. In turn, the 'green' label encoded as a 1 will be represented with a binary vector [0, 1] where the first index is marked with a value of 1.This idea can also be extended to a muliclass problem.

$$\begin{aligned} MSE &= \frac{1}{1}\sum_{i=1}^{1}\left(y_{true} - y_{pred}\right)^2 \\ &= \left(y_{true} - y_{pred}\right)^2 \\ &= \left(1 - y_{pred}\right)^2 \end{aligned}$$

Figure 6: MSE

F. Regularizers

Any neural network model show train just right and not over fit or under fit.Overfitting occurs when a model captures the noise of the data. Intuitively, overfitting occurs when the model or the algorithm fits the data too well whereas underfitting occurs when a model cannot capture the underlying trend of the data. Intuitively, underfitting occurs when the model or the algorithm does not fit the data well enough.It is shown below.Hence in order to avoid both the conditions, certain features called as the regularizers are used which is not but a technique which makes slight modifications to the learning algorithm such that the model generalizes better. This in turn improves the model's performance on the unseen data as well.Let us consider the regularizers used in this project in the following section.

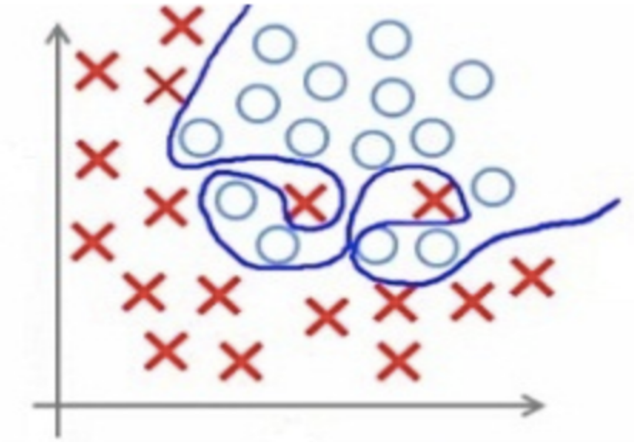


Figure 7: Over-fitting [2]

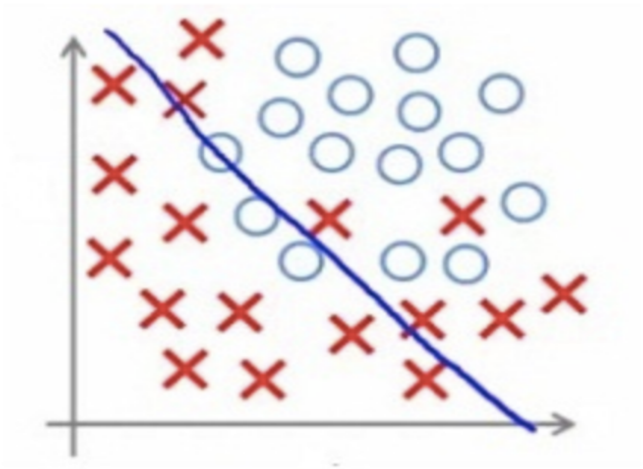


Figure 8: Underfitting [2]

G. Dropout

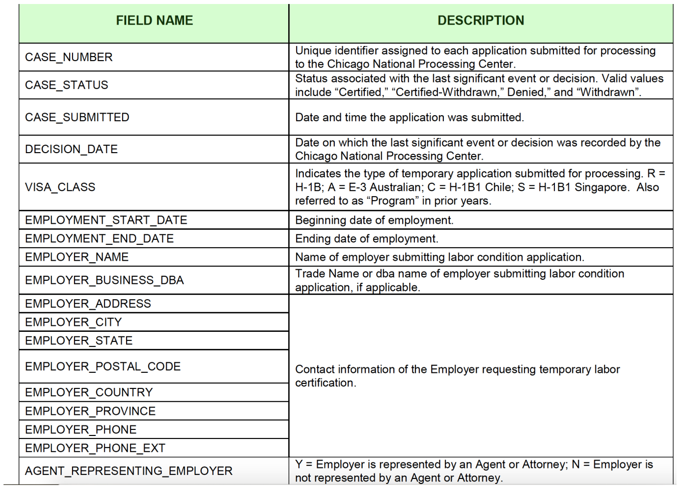
The most commonly used regularizer is dropout layer.Dropout is easily implemented by randomly selecting nodes to be dropped-out with a given probability (e.g. 20%) each weight update cycle.It can be applied to input neurons called the visible layer or hidden layers.The dropout rate is set to 20%,5 inputs will be randomly excluded from each update cycle.Dropout is used to avid over fitting of the model.

H. L2 and L1 regularization

L1 and L2 are the most common types of regularization. These update the general cost function by adding another term known as the regularization term.Due to the addition of this regularization term, the values of weight matrices decrease because it assumes that a neural network with smaller weight matrices leads to simpler models. Therefore, it will also reduce overfitting to quite an extent.Different types of kernel,bias,activity initializations are also applied while training the model.For example glorot normal is a type of kernel initializer.

IV. Dataset

The main dataset is downloaded from Kaggle listed under the name “H-1B Visa Petitions 2017“.It contains 53 features and 624650 records of applicants.The main step is pre-processing the data to exclude unnecessary features and also to convert every parameter into a number for training the neural network model.The output class label is converted into categorical and missing data if filled accordingly using mode values where ever necessary,to make the data uniform.Once the data processing was done, the data was divided into training ,testing set and validation sets.The data split was 33% testing and 10% validation.The description of the data is provided below in the picture.



FIELD NAME	DESCRIPTION
CASE_NUMBER	Unique identifier assigned to each application submitted for processing to the Chicago National Processing Center.
CASE_STATUS	Status associated with the last significant event or decision. Valid values include “Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”.
CASE_SUBMITTED	Date and time the application was submitted.
DECISION_DATE	Date on which the last significant event or decision was recorded by the Chicago National Processing Center.
VISA_CLASS	Indicates the type of temporary application submitted for processing. R = H-1B; A = E-3 Australian; C = H-1B1 Chile; S = H-1B1 Singapore. Also referred to as “Program” in prior years.
EMPLOYMENT_START_DATE	Beginning date of employment.
EMPLOYMENT_END_DATE	Ending date of employment.
EMPLOYER_NAME	Name of employer submitting labor condition application.
EMPLOYER_BUSINESS_DBA	Trade Name or dba name of employer submitting labor condition application, if applicable.
EMPLOYER_ADDRESS	Contact information of the Employer requesting temporary labor certification.
EMPLOYER_CITY	
EMPLOYER_STATE	
EMPLOYER_POSTAL_CODE	
EMPLOYER_COUNTRY	
EMPLOYER_PROVINCE	
EMPLOYER_PHONE	
EMPLOYER_PHONE_EXT	
AGENT_REPRESENTING_EMPLOYER	Y = Employer is represented by an Agent or Attorney; N = Employer is not represented by an Agent or Attorney.

Figure 9: Dataset description [4]

1) ***Pre-processing****:* The H-1B petitions data downloaded was not in proper format required for training a neural network model. Also, it contained junk data which had to pre-processed before proceeding further.The first step is to read the data as a pandas dataframe in python.Every column in the data represents the parameters which contribute for the final outcome.Hence, it is necessary to convert all the parameters accordingly and remove those which do not contribute.Accordingly out of 53 features, 27 were pre-processed and used to build the model. The steps followed for pre-processing them are as follows:

- **Case Number**:The first column represents the case number of an applicant.For example it is of the form I-200-16055-173457.Hence we preserve only the number part of it and replace other characters.

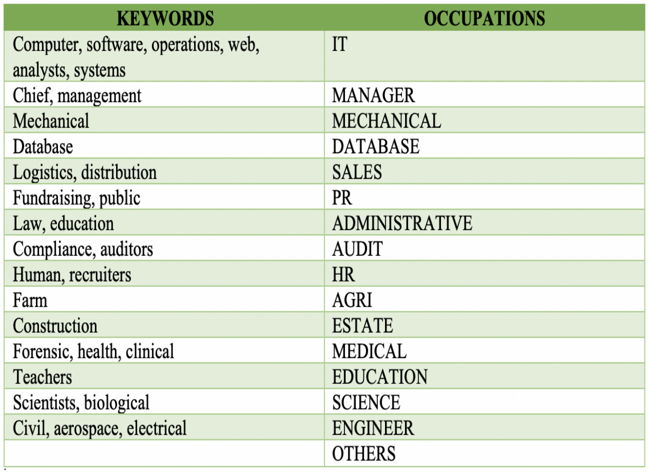
- **Case Submitted date, Employee Start date, Decision date**:and many more.The columns representing the data as a date was converted completely into seconds using pandas to date.

- **Employer Name**:This column represents the name of the company/institution seeking the visa.Since there were in textual format, it had to be converted into numerical.The idea is to set a threshold n the number of applications per company.The threshold was set to 500. Only when the employer had more than 500 petitions, I allocated a number to that.This way we can process all other companies with less applicants(since there were many with just 1 application).The data contained 72 such employers.The same procedure was followed even for processing job title of the applicant.The count was 151.All the missing values here, was filled with the mode value.

- **Occupation Name (SOC CODE)**:This column represents the occupational field of the applicant.SOC CODE is the Occupational code associated with the job being requested for temporary labor condition, as classified by the Standard Occupational Classification (SOC) System. The method followed to process this is by using keywords.If the occupation name contained the pre-defined keyword then it gets classified into a particular field.For example, if the occupation name contains the either one of the keywords computers,software or operations, the occupation of the person becomes IT.Similarly I used a pre-defined keyword set to classify them into 16 occupational fields.Finally, these 16 occupations were given a number that is each field gets a number from 0-15.The list of keywords and the occupations are tabulated below.

- **Wages**:There are mainly two types of wages called the prevailing wage and other source wage.First step is to convert all the wages in to yearly pay as the original data had all types of units of pay hourly,weekly,biweekly,monthly.I also retained the other source wages to study if it contributed to the model accuracy.

- **Wages**:There are mainly two types of wages called the prevailing wage and other source wage.First step is to convert all the wages in to yearly pay as the original data had all types of units of pay hourly,weekly,biweekly,monthly.I also retained the other source wages to study if it contributed to the model accuracy.There is also a column indicating the level of the wage(1-4) and also the different sources being OES, CBA, DBA, SCA.



KEYWORDS	OCCUPATIONS
Computer, software, operations, web, analysts, systems	IT
Chief, management	MANAGER
Mechanical	MECHANICAL
Database	DATABASE
Logistics, distribution	SALES
Fundraising, public	PR
Law, education	ADMINISTRATIVE
Compliance, auditors	AUDIT
Human, recruiters	HR
Farm	AGRI
Construction	ESTATE
Forensic, health, clinical	MEDICAL
Teachers	EDUCATION
Scientists, biological	SCIENCE
Civil, aerospace, electrical	ENGINEER
	OTHERS

Figure 10: Keywords and Occupations

- **Output Label**:The Case Status becomes the output class label for this supervised learning problem.This parameter indicates the outcome of the applicant's petition indicating the eligibility.It can be of 4 types:certified which indicate that the person is eligible to file for the visa,denied indicates that the person is not eligible for the visa.The other two are certified-withdrawn and withdrawn.Withdrawn represents the condition where the applicant withdrew his petition and certified-withdrawn represents the condition where the applicant is n longer working with the employer under whom he/she had received the visa.

  I decided to study this as a multi-class problem retaining all four class labels because the certified-withdrawn did not have many entries and instead of converting the withdrawn as denied , I studied it as a separate class only to study if it affects the performance of the model. The problem could also be studied as a binary classification problem by using just denied and certified and necessarily eliminating other class labels.

*2) **Final Features**:* This section summarizes the final 27 features and their descriptions which were considered while training the model.

- **Case Number**:This feature represents the Unique identifier assigned to each application submitted for processing to the Chicago National Processing Center.

- **Case Submitted**:This feature represents the date and time the application was submitted.

- **Decision Date**:This feature represents the date on which the last significant event or decision was recorded by the Chicago National Processing Center.

- **Employment Start Date**:This feature represents the Beginning date of employment.
- **Employment End Date**:This feature represents the end date of employment.
- **Employer Name**:This feature represents the name of employer submitting labor condition application.
- **Agent Representing Employer**:This feature represents the if Y = Employer is represented by an Agent or Attorney; N = Employer is not represented by an Agent or Attorney.They were converted to 0,1 respectively.
- **Job Title**:This feature represents the Title of the job.
- **Total Workers**:This feature represents the total number of foreign workers requested by the Employer(s).
- **New Employment**:This feature indicates requested worker(s) will begin employment for new employer,as defined by USCIS I-29
- **Continued Employment**:Indicates requested worker(s) will be continuing employment with same employer, as defined by USCIS I-29.
- **Change Previous Employment**:Indicates requested worker(s) will be continuing employment with same employer without material change to job duties, as defined by USCIS I-29
- **New Concurrent Employment**:Indicates requested worker(s) will begin employment with additional employer, as defined by USCIS I-29.
- **Change Employer**:Indicates requested worker(s) will begin employment for new employer,using the same classification currently held, as defined by USCIS I-29.
- **Amended Petitions**:Indicates requested worker(s) will be continuing employment with same employer with material change to job duties, as defined by USCIS I-29.
- **Full Time Position**:This feature indicates if Y = Full Time Position; N = Part Time Position.They were converted to 0,1 respectively.
- **Prevailing Wage**: This feature indicates the prevailing Wage for the job being requested for temporary labor condition.
- **Wage Source**: This feature indicates the variables included OES, CBA, DBA, SCA or Other.
- **Wage Level**: This feature indicates the variables include I, II, III, IV or N/A.
- **Wage Source Year**: This feature indicates the year the Prevailing Wage Source was Issued.
- **Wage Source other**: This feature provides the source of wage if there is other source of income.
- **Wage Rate**:This feature indicates the employer's proposed wage rate.
- **H-1B Dependent**:This feature indicates the dependability.If yes,or else no.Y = Employer is H-1B Dependent; N = Employer is not H-1B Dependent.They were converted to 0,1 respectively.
- **Willful Violator**:This feature indicates the dependability.If yes,or else no.Y = Employer has been previously found to be a Willful Violator; N =Employer has not been considered a Willful Violator.They were converted to 0,1 respectively.
- **Support H-1B**:This feature indicates the dependability on the applicant.If yes,or else no .Y = Employer will use the temporary labor condition application only to support H-1B petitions or extensions of status of exempt H-1B worker(s);N = Employer will not use the temporary labor condition application to support H-1B petitions or extensions of status for exempt H-1B worker(s);They were converted to 0,1 respectively.
- **Labor Condition Agree**:This feature indicates the dependability on the applicant.If yes,or else no.Y = Employer agrees to the responses to the Labor Condition Statements as in the subsection; N = Employer does not agree to the responses to the Labor Conditions Statements in the subsection.They were converted to 0,1 respectively.
- **Original Certificate Date**:This feature indicates the original Certification Date for a Certified Withdrawn application.
- **SOC Name**:This feature indicates the occupational name associated with the SOC CODE.
- **Case Status**:This feature indicates the status associated with the last significant event or decision.

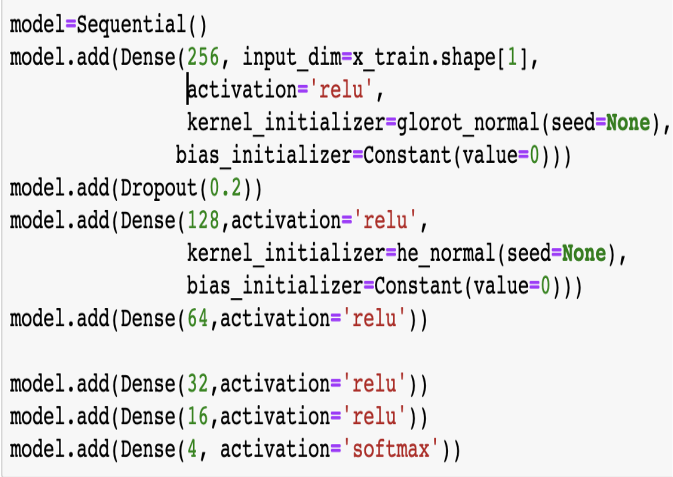
The other parameters were dropped.The names of the dropped parameters are: unnamed: 0,visa class,employer business dba,employer country,employer province,employer phone extension,employer address,employer city,employer state,employer postal code,agent attorney name,agent attorney city,agent attorney state,employer phone,soc code,naics code,public disclosure location,worksite city,worksite county,worksite state,worksite postal code,pw source other,wage rate of pay from,wage unit of pay.

V. Method

It involves constructing a sequential neural network model with the required regularizers to achieve better performance.The number of neurons in each layer, number of layers to be considered by chosen by trial and error.To understand and study the models better, five sequential models were built with different kinds of regularizers, loss functions and activation functions.The final goal was to actual the average accuracy's across them to cross verify the learning of the data.

- **Sequential model**: Sequential model using fully connected layers was used for construction.The fully connected layer consists of fully connected neurons, meaning each neuron is dependent on all the output of the previous layer.Fully connected layer was used as the data considered is independent of one another and also does not consist of images.
- **Training**:Each model was trained with different parameters for 10-20 epochs.The validation was done on 10% of the data where as the learning rate was chosen differently for different models.The output layer activation function was Softmax where as different models different activation functions in different layers.Dropout was used for both visible layer as well as hidden layer.
- **Experimentation**:In order to calculate the performance of the model generated, different forms of metrics are used.The popular ones among them as scatter plot, co-relation matrix, accuracy plots and confusion matrix.For our work,let us stick to plotting accuracy and loss function along with confusion matrix.

The figure 11 represents the sample of a model and its parameters.

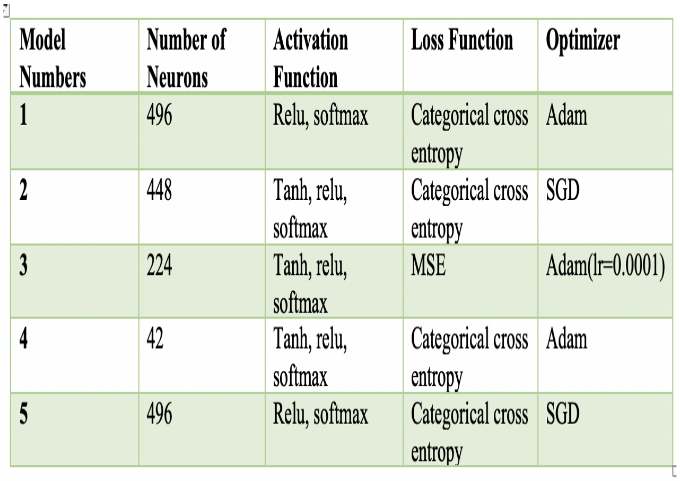


```
model=Sequential()
model.add(Dense(256, input_dim=x_train.shape[1],
                activation='relu',
                kernel_initializer=glorot_normal(seed=None),
               bias_initializer=Constant(value=0)))
model.add(Dropout(0.2))
model.add(Dense(128,activation='relu',
                kernel_initializer=he_normal(seed=None),
                bias_initializer=Constant(value=0)))
model.add(Dense(64,activation='relu'))

model.add(Dense(32,activation='relu'))
model.add(Dense(16,activation='relu'))
model.add(Dense(4, activation='softmax'))
```

Figure 11: Sequential Model

The table below gives an elaborate representations of the number of parameters used and also provides the different types of activation functions used along with different loss functions.



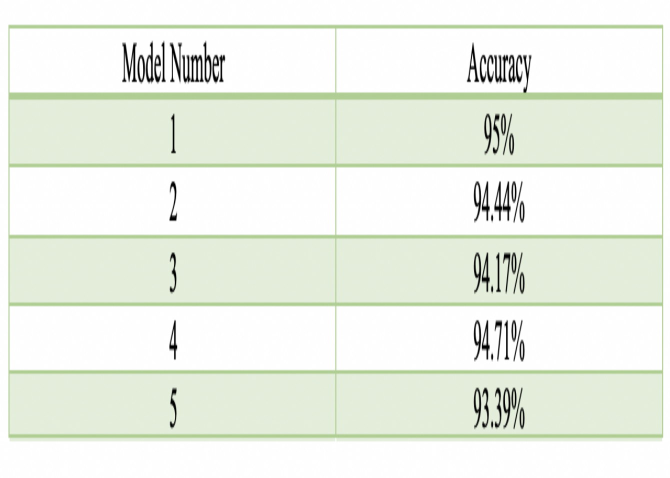
Model Numbers	Number of Neurons	Activation Function	Loss Function	Optimizer
1	496	Relu, softmax	Categorical cross entropy	Adam
2	448	Tanh, relu, softmax	Categorical cross entropy	SGD
3	224	Tanh, relu, softmax	MSE	Adam(lr=0.0001)
4	42	Tanh, relu, softmax	Categorical cross entropy	Adam
5	496	Relu, softmax	Categorical cross entropy	SGD

Figure 12: Model Parameters

VI. Results

A. Accuracy

The table below provides the accuracy of all the five models and the next section provides the loss as well as accuracy graphs for both training and validation sets.



Model Number	Accuracy
1	95%
2	94.44%
3	94.17%
4	94.71%
5	93.39%

Figure 13: Model Accuracy

B. Confusion Matrix

A confusion matrix is a summary of prediction results on a classification problem.The number of correct and incorrect predictions are summarized with count values and broken down by each class. Expected down the side: Each row of the matrix corresponds to a predicted class.
Predicted across the top: Each column of the matrix corresponds to an actual class. The total number of correct predictions for a class go into the expected row for that class value and the predicted column for that class value and the way for incorrect as well.

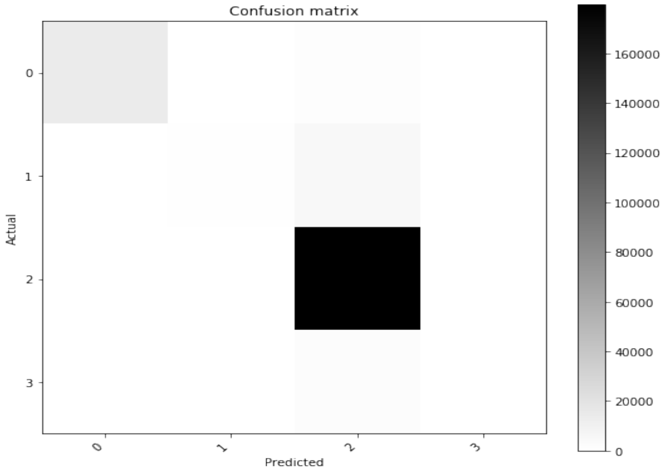


Figure 14: Confusion Matrix

C. Plots of Accuracy and Loss

The plots represent the training as well as validation accuracy and the same applies to loss as well.We can see in the plots that the validation loss decreases or sometimes remains the same.

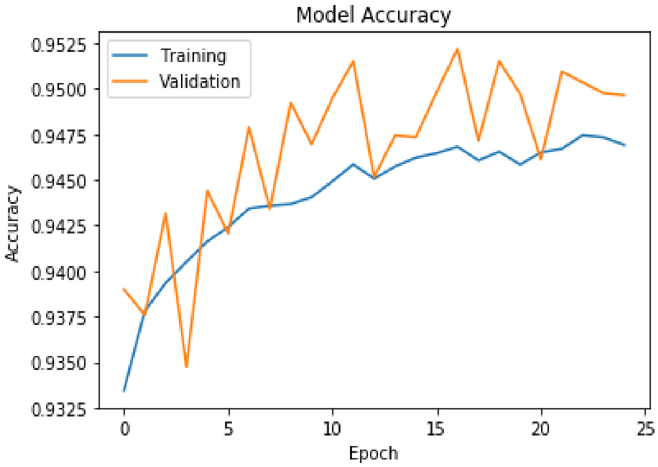


Figure 15: Model1 Accuracy

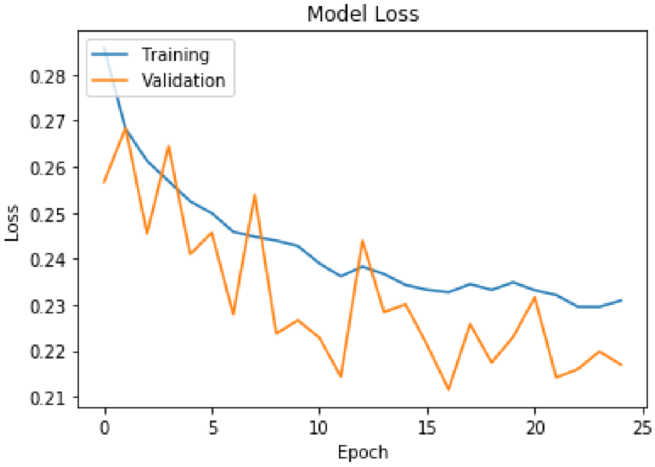


Figure 16: Model1 Loss

VII. Conclusion

The current work focused on predicting the eligibility outcome of H-1B visa petitions.The main idea is to train a neural network model with dense layers and predict the outcome.Along the process it made use of necessary regularizers,activation functions,loss functions to train the model right and not allow it to over or under fit.Different activations were used in the input and hidden layers and softmax function was used in the output layer since the work trained the model as a multi-class classification problem.I first converted the output class label as a categorical data using the one hot encoding method.The accuracies are almost as high as 94%,hence it it safe to say that average overall model accuracy is 94%.

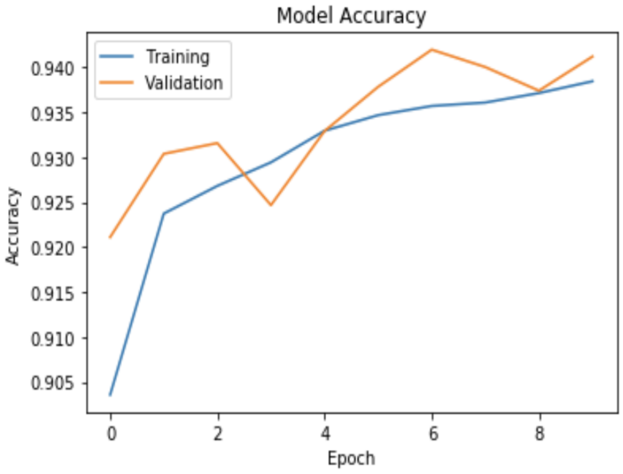


Figure 17: Model2 Accuracy

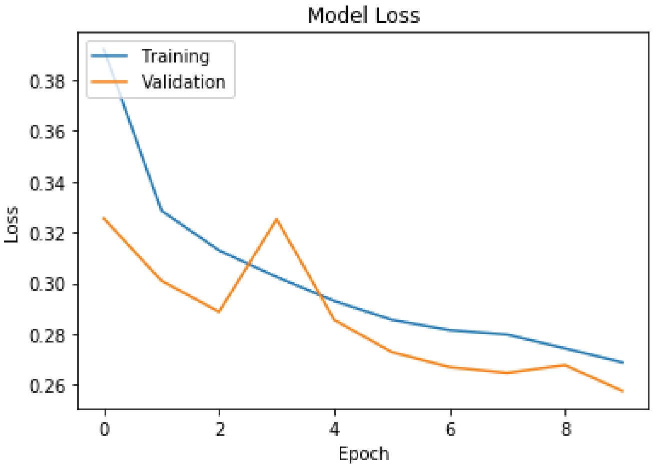


Figure 18: Model2 Loss

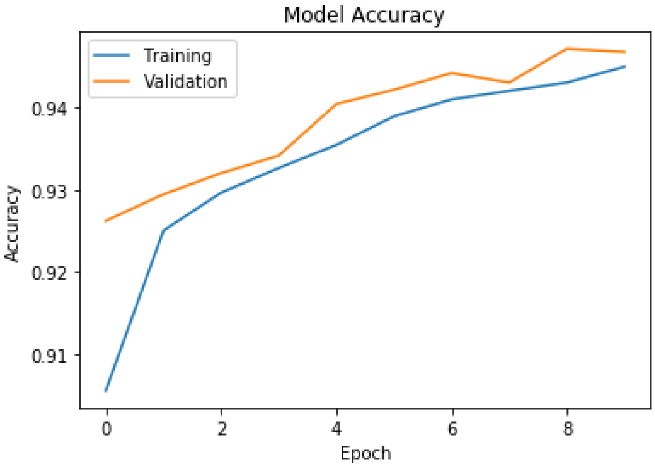


Figure 19: Model3 Accuracy

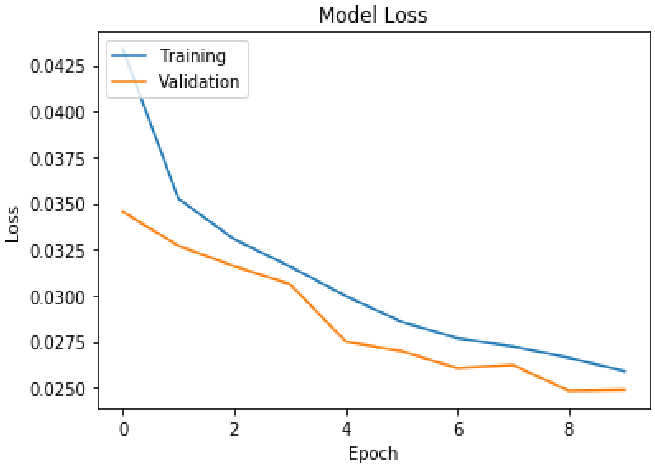


Figure 20: Model3 Loss

VIII. Future Work

For future work we can consider the entire work as a binary classification problem by using just denied and certified case status as the class label.This way it is possible to train and model different models and apply voting classifier or voting ensemble model.There is also room to consider different parameters while training the model.One more idea could be to use GRU units in the sequential model.There is a huge scope to apply the same on different data as well.There is also a huge scope for performing different types of clustering when approached as an unsupervised learning problem.

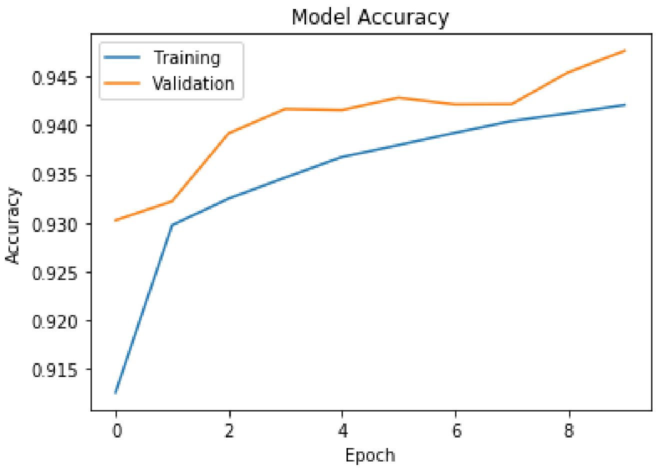


Figure 21: Model4 Accuracy

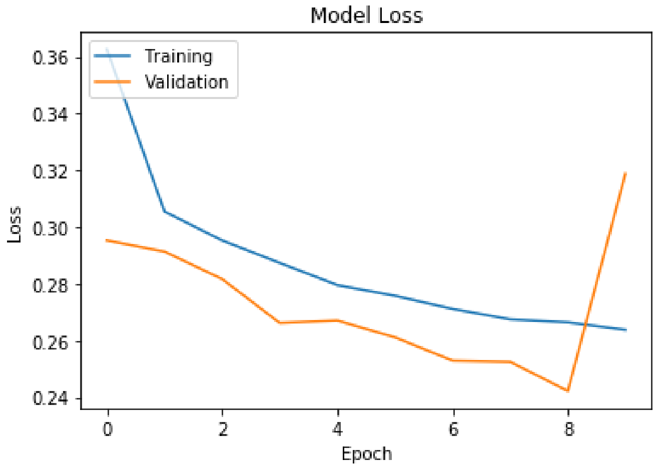


Figure 22: Model4 Loss

References

[1] www.google.com

[2] www.analyticsvidhya.com

[3] www.machinelearningmastery.com

[4] https://www.kaggle.com/jonamjar/h1b-data-set-2017

[5] https://www.datacamp.com/community/tutorials/predicting-H-1B-visa-status-python

[6] https://www.kaggle.com/dpandya18/h1b-visa-status-prediction

[7] https://datascience.stackexchange.com/questions/40067/confusion-matrix-three-classes-python

[8] https://www.geeksforgeeks.org/multiclass-classification-using-scikit-learn/

[9] Multilayer Perceptron Neural Network (MLPs)For Analyzing the Properties of Jordan Oil Shale,

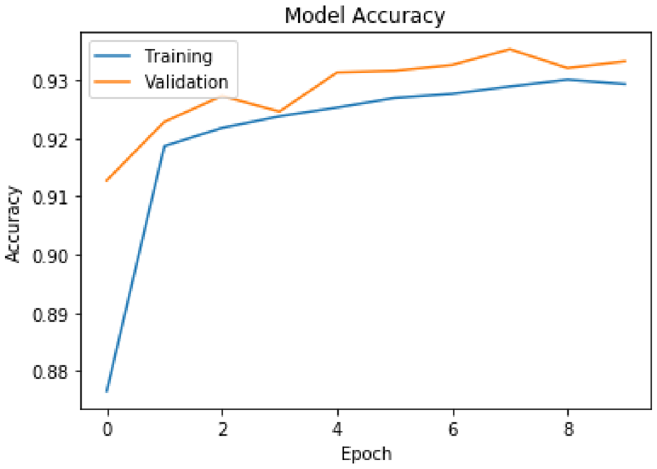


Figure 23: Model5 Accuracy

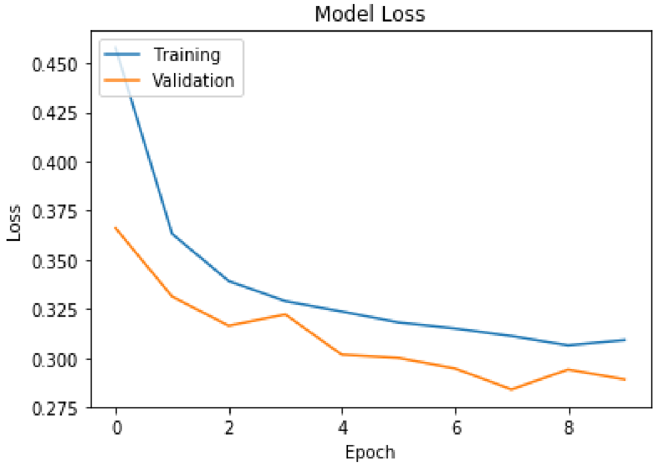


Figure 24: Model5 Loss