Census Analysis with National Economic Projections using Advanced Data Mining

Abstract— **The power of healthy economy is uncanny. For any given country its development depends on the people and the working infrastructure that they imbibe with. Economic growth plays a vital role when we comment on development in terms of financial growth. This paper addresses the National Economic Projections and forecasts their repercussions. For an overall development it is important to keep a track of these figures. The concerned paper serves the purpose of analyzing Population Census and then adheres its consequences on Economic National projections. The prime focus of it is to forecast the census and then relate these predictions with historical Employment and Unemployment data. On implementation grounds, the Census rate is predicted and further to which the Employment and Unemployment rate is specifically considered and the analysis is performed in two most influential domains of Big data, that is Machine Learning and Deep Learning. The observed results appear to be robust and the models have performed to its fullest giving the most optimal performance parameters. The amalgamation of Census data alongside Employment and Unemployment rate with the ken of forecasting turns out to be the novelty of the paper.**

Keywords—Random Forest, KNN, Ensemble, RNN, ARIMA, C5.0, Decision Tree.

I. INTRODUCTION

The Census prediction terminology is been a hype for quite a long time. The reason behind these analyses is the fact that the Census data is fairly numerical in nature and it servers many challenges in both statistical and economic domain. There is some decent amount of work done when it comes to predicting the census rate as it helps in maintaining the track of population changes. These predictions give certain rough figures and contributes in saving a lot of efforts that come into account while collecting the census
data. Certain work contributions are cited in the below section, though the work done is recommendable but majority of the analysis lies in the statistical domain and machine learning appears to be a future scope to these implementations. This challenge is identified by this paper and hence the projections appears to be undertaken by using Ensemble techniques and Neural networks.

To start on with, the project narrowed down its analysis for Census data of United states of America. The population estimates from year 2011 were taken, the data here was a times series data recorded on monthly basis. The most efficient Time series algorithm which is ARIMA, is been used to predict the 2018 population. The obtained result gave a fairly increase in the observed records. These analyses were further considered in order to comment on the Employment and Unemployment rate. For the Census effects on unemployment, the census data alongside unemployment rate for counties in Unites states is analyzed. The data is fully numerical, that is converted to categorical and later merged. The results comment on the unemployment rate class in which each county lies. The most advanced method of Machine learning i.e. Ensemble is been used to undertake these performance parameters. For starters the feature engineering was done by using Random forest. The ensemble was built later based upon these features using three dominating numeric data models which are KNN, C5.0 and Decision Tree. Further to these the Census effects on Employment rate were observed by one of the most meticulous time series evaluation models i.e. Recurrent Neural Networks (RNN). These implementations efficiently drag the project from a machine learning domain into a Deep learning protocol. The outcome of RNN effectively predict the testing data results and models appears to be a best fit and is tuned with decent efficiency.

Hence, the given paper is a first ever implementation that amalgamates Census data with Employment and Unemployment rates, comments on its consequences on the same and all of these is productively carried out in Machine learning and Deep learning field. The data is trained, tested and analyzed in time series and classification domain. The implementation appears to be a Novel approach that adhered the proposed research question 'How can we analyse the effects of Census prediction on Economic national Projections?' and adequately answered it with highly promising models that performed fairly decent.

II. RELATED WORK

As mentioned earlier there is a decent amount of work done in the prediction of Census and the Economic projections. Several platforms and domains like statistics, web clouds and real time economic data storage are few examples. One such related work encorporates web technology, which is used to predict unemployment by extracting economic indicators by Mioara [1]. Google trend tool was used for real time data source than the traditional method of collecting the data and publishing it by the government agencies or by public institutions. By adjusting Autoregression (AR) model the forecasting accuracy was improved was seen from the results [1]. Another developed implementation for population prediction is given by Yunong Zhang in [8], where in the use Chebyshev-activation WASD (Weight and Structure Determination) neuronet approach was done to obtain the

population census prediction. Performance of neuronet is dependent on hidden neuron layers, a smaller number of layers may not give the desired output whereas a greater number of hidden layer neurons may result into overfitting. WASD neuronet provides better results than traditional BP neuronet. Population can be predicted by three-layer feed forward neuronet equipped with WASD algorithm. Thus, with 69.7% of total factors show that there will be decline trend in European population in next decade [8].

In [2], which is considered as a base literature with regards to our model implementation where in a case study of population with Fisheries is done on basis of Ensemble modelling. Sean C Anderson proposed that average population prediction of several models is taken to gain more accurate results. Results of individual models may give uncertainty and errors may occur, therefore ensemble methods are preferred which reduces the chances of errors. Initially, individual assessment models are estimated, and later cross validation of built dataset is simulated. Simulated datasets against a parameter is then tested. In this paper, global fish stocks are considered against the simulated dataset. Individual models are initially fitted to training data. Estimations from these models are then combined based on the covariates. On simulated dataset of fisheries ensemble methods are tested with the help of FLR (FLBRP) packages for statistical software R. Time Series parameter and fishery characteristics are considered in framework to generate population projection and resulting catch time series [2].

Edgar Morgenroth, in [3] described importance of census prediction which is related to various factors such as public services provision, land use planning. Evaluation of different techniques and methodology of forecasting performance for period of 1991-1996 is performed. Cohort component method, simple extrapolation, regression-based extrapolation, corelated indicators are the different methods used to generate population projections. According to the results obtained extrapolation techniques perform well as compared to cohort component model [3]. This paper majorly focuses on ARIMA and cohort component and Advances data mining techniques for big data are considered as a future scope for the literature, which is taken care of in our implementated paper.

On the pararell grounds with population projection, Dinesh Gawatre in [4], explained that various factors are dependent on population, so prediction of population is required to provide different facilities to public. Arithmetic increase method, geometric increase method, incremental increase method, Simple graphical method and logistic curve method are carried for accurate forecasting of population. From obtained graphs it is concluded that geometric increase method tends to give higher accuracy and behaves exponentially. The implementation is purely in statistical domain of accuracy and again machine learning is the future scope for the given paper. Another statistical implementation proposed by U. Mallick and H.Biswas [5]developed a mathematical model which helped in reduction of unemployment through optimal control strategies where Pontryagin's Maximum Principle was used which establish two control strategies. Two control strategies give better result which increases employed population than the policies applied for employment by government [5]. These two different litratures work on the estimations of Census, Employement and Unemployement but majorly on statistics domain and are dealth seperatly. Hence, our implemaentaions not only combines these three parameters but undertakes its performance with effect of advanced data mining techniques of machine learning and data mining.

For the implementation in time series ARIMA model evaluation is done in this paper. The supporting literature for this implementation was studied from [10], where the authore described process for analyzing the subset order in Autoregression Integrated Moving Average (ARIMA) on the basis of overfitting concept. Indonesia's inflation dataset is used in the proposed paper. Order of AR, MA is identified on the basis of ACF model. Accurate order of subset is obtained from ARIMA model for Indonesia's inflation dataset. AR tails off exponential decay and cuts off after lag p whereas, MA cuts off after lag q and tails off exponential decay. tail off in ACF occurs after lag(q-p) in ARMA whereas tails off occur after lag(p-q) in ARMA. Parameters of the model were evaluated by various methods such as Maximum Likelihood (ML) method, Unconditional Least Squares (ULS), or Conditional Least Squares (CLS) method. Ljung-Box test is considered to identify whether assumptions of independence of residuals are satisfied or not. Order of subset ARIMA subset can be taken into consideration to reduce RMSE, AIC and SBC and best mean of ARIMA model can be used for forecasting [10]. These concepts are well versed while implementation and the results of ARIMA were evaluated on basis of the same.

Another study that took ARIMA into account was by Wiwik Anggraeni [6], author forecasted the demand for clothes based on the Eid holidays. ARIMA and Arimax methods are compared for forecasting the demand. To increase the accuracy of ARIMA model the number of variables is increased. AIC, MAPE, and RMSE values for ARIMA are 929.693, 17.29% and 73.56 whereas for ARIMAX it is 891.144, 14.80% and 72.09 respectively. Thus, comparison gives that Arimax model gives better performance than ARIMA [6]. Considering this a similar study by Yanming Yang, suggested Arima model to forecast Aero-Material consumption which shows seasonal change. The Winter model and Seasonal ARIMA models are used to compare the forecast. MAPE, MAD, and MSD are 0.22712, 0.00795 and 0.00075 respectively and for Winter Model it is 0.34456, 0.17511 and 5.81527 respectively. The model fits better for small values. Thus, from the results of prediction, seasonal ARIMA can accurately predict and provide the aero-material consumption rate [7]. By considering these studies the decision of evaluating ARIMA as a best suited alogorithm for this paper was made.

The literature by Kinley Wangdi in [9], appears to be efficient for time series analysis. The author also approached ARIMA to performed prediction and forecasting of malaria incidence in an endemic area of Bhutan. Dataset from 1994 to 2008 was used to forecast 2009 and 2010 malaria cases. Dataset was checked whether time series was stationary or non-stationary. Feasible models were identified based on ACF (Autocorrelation functions) and PACF (Partial autocorrelation functions). Arima model of time series analysis was useful to forecast endemic areas of Bhutan. The best fit model of ARIMA (1,1,0) (0,1,1). The error percentage obtained was -8.12%. [9].

The deep learning model was also developed for predicting the value of bitcoin. As mentioned in [11], LSTM model is used in time series problems. The performance of CPU and GPU was tested with respect to machine learning models to predict the bitcoin value. The ARIMA model was implemented as a base model which was outperformed by LSTM model with the accuracy of 52%. The other paper focuses on dealing the multivariate time series data with missing values using GRU RNN model [12]. The RNN was used in [13] to predict the rainfall in Bihar region to predict the flood-based deaths of human and animals. The results achieved were not so accurate as the variables were not descriptive.

III. DATA MINING METHODOLOGY

For an efficient implementation with desired results the research project must follow a data mining methodology which is best suited to solve the problem. There are multiple methodologies used for implementing data mining models based on the criteria, resources and issue to be solved. Some of the widely used data mining methodologies are KDD, CRISP-DM and SEMMA approach. The KDD approach follows certain steps to analyze and extract knowledge from the data. The steps performed in KDD approach are as follows:

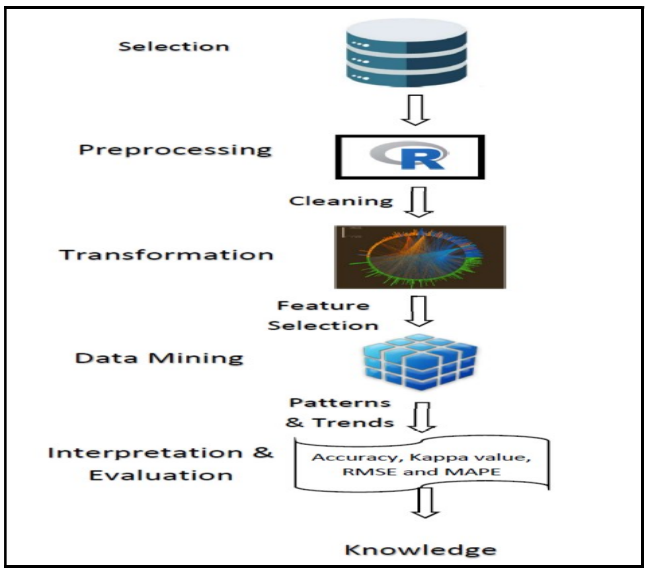

Figure 1: KDD Methodology

The KDD approach is basically used by marketing, finance, investment and other business application areas. All these application areas are time oriented and maintains a time series. This paper also depicts a time series problem and needs to follow the KDD approach for knowledge discovery.

A. Data Selection

The KDD approach starts with the data selection phase which aims to select the appropriate data from web pages in order to answer the research question and fulfil the desired objectives. The data selection is a crucial phase in a data mining methodology as the attributes mentioned in the datasets must be closely associated and co-related with each other to achieve better results. The datasets are collected from two different sources:

1. Data.gov:

   This website is the repository for US government data. We have collected two datasets from this repository which represents US population and unemployment data. Both the datasets are uploaded on 4th Feb 2018. It represents the population and unemployment metrics of the US country with respect to each county of US. There are nearly 3100 rows and more than 100 columns [14] [15].

2. Fred Economic Data:

   This website provides economical data of US county. The data provided helps in employment prediction and population forecasting as the data shown here is in seasonal format and can be used for time series models [16][17].

B. Preprocessing

The datasets downloaded are not in a required format to process and thus pre-processing of the data has to be done before proceeding further. Once the data is cleaned and pre-processed as per the required format, we can implement various models to accomplish the research objective.

Pre-processing consists of various tasks such as removing unnecessary attributes, outliers, null values, missing values noisy data, etc. Each cleaning task was done in Rstudio using various inbuilt packages. The pre-processing done in this research is explained below:

- Removed meta-data descriptor from the datasets using Rstudio.
- Two different datasets were merged together with the help of FIPS column which was common in both the datasets.
- Removed all missing values and NA values from the combined dataset as the count of missing values were high and imputing methods can lead to wrong analysis and prediction.

- The columns such as population consisting of comma separated values i.e. 3,51,000 were cleaned by eliminating commas i.e. 351000.
- Typecasting was performed to convert the datatype of the data from factor to numeric as it is essential for implementing data mining models.
- The dataset values were scaled for performing KNN model as it is a requirement for implementing KNN.
- As our target variable (unemployment rate) was continuous, we converted the values into categorical form using cut and break function. We distributed the data equally with the help of table function to avoid class imbalance.

C. Transformation

Once the data is cleaned and pre-processed it is ready for transformation phase. This phase focuses on projecting important features to be selected while implementing and analysing the suitable model for the dataset. Basically, useful features were selected based on the objectives of the research project. We performed random forest algorithm on the data as a feature selection technique for ensemble model. The attributes returned by accuracy and ginni plot of random forest were used for ensemble modelling as it identifies the best contributors for the model to gain better accuracy and fitness of the model.

D. Data Mining:

One of the crucial tasks in data mining is analyzing the data and implementing a model which provides better solution to the given problem statement. Once the models are finalized and implemented, we need to evaluate the results of those models to identify which model is best suited for predicting the population and analyzing the unemployment rate and employment rate of United States of America.

Before proceeding to the implementation phase, we need to divide the datasets into training and testing data. The training data holds on to 75% of the data to identify the hidden patterns whereas testing data carries 25% of the data.

To begin with the modelling process, we need to identify the objective of this research. The aim to predict the population of USA in 2018 can be done with time series model whereas analyzing the unemployment rate and employment rate of USA is a classification problem and needs to be solved using classification models. Thus, the research consists of time series models such as ARIMA and RNN for population and employment rate prediction. The research also consist classification model such as Ensemble for analyzing unemployment rate of US. The models implemented are discussed below:

1. ARIMA Model

The ARIMA model appears to be an appropriate selection when it comes to analysis of the time series data. The first data source [16], where the population figures are recorded on monthly basis. The data set was imported in R studio and encoded with auto.arima function using libraries like: MASS, tseries and forecast that resulted in the prediction of the population of 2018.

The .csv data was converted to time series and then lagged with frequency of 12. The frequency selection was done as 12 because the data is ranged over monthly period for each year. After which auto correlation and partial correlation was injected to the time series data. The reason for doing this is to obtain the correlation/connection of the given data with its past values as these parameters would assist primarily on obtaining the forecast results. Figure 2: indicates the auto correlation output. Where in each line that goes in upward direction represents the correlation between the lags.

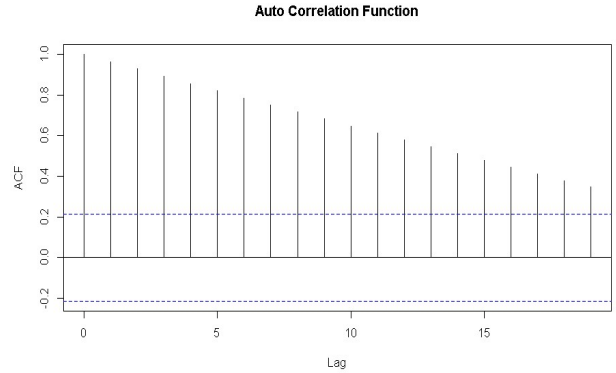

Figure 2: Auto Correlation

Followed to this figure 3 indicates the result of partial correlation performed. According to the output obtained we can observe a sudden decrease in the lag values. This sudden decrease indicates that the data is stationary. For time series analysis we desire to have a stationary data as it comes up with constant mean and variance values.

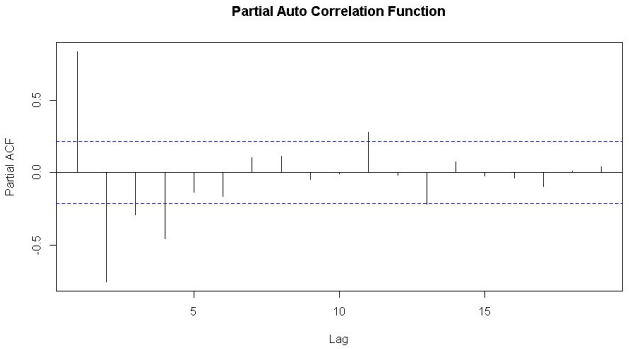

Figure 3: Partial Correlation

On the similar grounds to check the stationarity of the data one of the most promising Statistical test called as the Dickey Fuller test of hypotheses is carried out. As per which we consider the null hypothesis as the data not to be in stationary form and then observe the results. After applying the test to our dataset the observed significance value is given in the figure below.

```
> adf.test(lnpop) #by DF test p-value <.05 indicates that we reject the null hypothesis of non stationarity.

        Augmented Dickey-Fuller Test

data:  lnpop
Dickey-Fuller = -1.3863, Lag order = 4, p-value = 0.8276
alternative hypothesis: stationary

> adf.test(diffpd)

        Augmented Dickey-Fuller Test

data:  diffpd
Dickey-Fuller = -7.1993, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Figure: 4 Dickey Fuller Test results

As per the results the lag order of 4 was set after the test which is a decent order number. When the test was done on the given data the significance value (p-value) appeared to be on a higher level as compared with the test result after differentiating the dataset. The importance of differentiation is explained in figure 5, as of now considering the significance value for the differentiated data, as per the Dickey fuller test if the P-value <0.05 we reject the Null hypothesis. Hence as per our results p-value = 0.01 (0.01<0.05) as a result of which the estimates that the data is stationary are confirmed after rejection of Null hypothesis.

Consider the differentiation of data, the order of differentiation was obtained by the 'ndiff' function in R studio that gave the results as in Figure 4 as:

```
> ndiffs(pd$Pop)
[1] 1
```

Figure 4: Selection of differentiation

From this the [p,d,q] and [P,D,Q] arima performance parameters were set. These values play a very important role in arima implementation as they set the forecast order and the integration of time series data. Where [p] resembles the Auto correlation of the data points, the [d] resembles the integration of the values and the [q] decides the moving averaged of the given dataset. For our data values an matrix of (3,1,2)(0,0,2) with frequency [12] is observed and the forecast results are as figure 6.

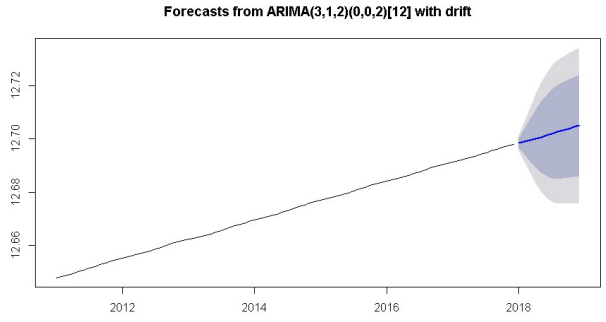

Figure 5: ARIMA result: Population 2018 projection.

After differentiation of the lagged time series data the auto.arima result is given above where the forecast is in correlation with historic data and the blue colored line indicates an increasing inclination of the population figures. In terms of prediction figures the same can be seen as given in below figure 6 where we have Actual data and the predicted data.

Actual Population	Forecasted Population
311078	327273.2
311234	327441.0
311400	327616.4
311575	327800.3
311757	327992.3
311956	328191.8
312168	328401.7
312388	328617.1
312612	328833.0
312817	329042.5

Figure 6: ARIMA forecast values

As a summary of the obtained data (Figure 7) we observe the RMSE value, this value with a minimum figure is desired because as lesser the figure better is the model fit. In our case RMSE= 0.00138 which is a considerably a low number that indicates our model to be a better fit with an equivalent low MAPE of 0.0128 indicating lesser error rate.

```
Poparima <- ts(lnpop, start = c(2011), frequency= 12)
fitpd <- auto.arima(Poparima)
fitpd
summary(fitpd)
                      ME        RMSE         MAE          MPE        MAPE        MASE         ACF1
Training set 0.0001507118 0.001380007 0.0001621431 0.001191601 0.001281832 0.02235039 -0.005332354
```

Figure 7: ARIMA model summary

2. RNN Model

After analyzing the results of the ARIMA base model to forecast the US population for 2018, we switched to neural network for predicting the employment rate. RNN (Recurrent Neural Network) model was evaluated for this research question as it is the updated model in the current era which aims to give the best results for the sequential data. Sequential data is directly related to time series data and the dataset used for this research question is also listed in the series of time. LSTM (Long Short-Term Memory) is an RNN based neural network model used for this research project as it provides an effective solution for the time series questions.

Before implementing RNN on dataset [17], normalization of the data was done as a preprocessing task for deep learning model. The input data was also reshaped according to the RNN requirement. After normalizing the dataset, the data was divided into training and testing dataset to check the accuracy of the trained RNN model. The trained dataset was based on the data containing employment rate of the previous 90 months of United States.

The model was developed in python programming with the help of keras [18] library in spyder framework. While building RNN model, we added dropout to avoid overfitting of the model. Dropout is a method used by LSTM units to exclude input and recurrent network from

weight updates during model training. The keras library was imported for estimating the dropouts in RNN model. These dropout regularizations were further evaluated with 4 layers to pass the neuros of 150 unit for each layer. This concept helps to increase the model performance and train model efficiently.

The other parameter to look into is evaluation of best optimizer for RNN model. The proposed RNN model uses 'Adam' optimizer for compiling the model as it carries the advantages of both AdaGrad and RMSProp optimizers which runs on the basis of moving average technique which is suitable for time series problems.

Once the compiling process was done, the recurrent neural network was trained on the data with epoch and batch size parameters. When the dataset enters the neural network and is passed in both forward and backward direction once then it is said to be one epoch. We have tested the RNN model with different epoch parameter and 120 was the best value suited for our prediction. After setting all the important parameters, the RNN model was trained successfully and tested against the test data to predict and visualize the output.

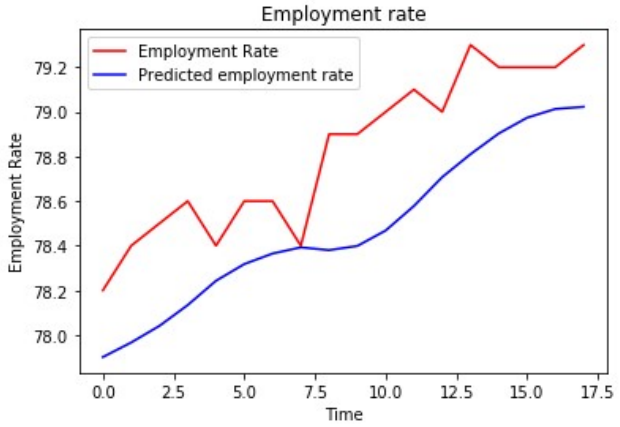

Figure 8: RNN forecast

The above diagram shows the trend recognized by the RNN model from the actual data. The actual employment rate was raised from 78.50 to 73.25 and the predicted values were close to actual figures showing trend in employment rate from 77.50 to 78.25.

```
    ...: rmse
Out[248]: 0.37405746987238236

In [249]: mse = mean_squared_error(ptest, predicted_Emp)

In [250]: mse
Out[250]: 0.13991899076732822
```

Figure 9: RNN RMSE result

The RMSE value of lower threshold was observed that indicates best fit of the given model. As shown in the above figure, the RMSE value is 0.37 which indicates that the model is performing well.

3. Ensemble Model

Ensemble is a predictive modelling technique which combines two or more models into one base model to increase the accuracy of the individually used models. The ensemble model used in this paper uses simple averaging and stacking ensemble techniques. The ensemble model comprises of three classifiers namely KNN, C5.0 and Rpart. Each classifier is explained below with respect to different ensemble techniques.

We have implemented C5.0, KNN and Rpart as our base learners for ensemble model. Here, C5.0 was used as it works on most of the classification problems and considers only important features for model building. It also performs well with small number of rows. Rpart acts like a decision tree and follows divide and conquer approach for predicting the result. KNN model was used as it is robust in nature and needs scaled values. The KNN model gives the accuracy of 70.81% with kappa value of 0.608. The accuracy calculated by c5.0 classifier is 72% with kappa values of 0.62 which is highest amongst all individual models. Rpart is the third model used in ensemble with accuracy of 71% with kappa value of 0.62 which performed slightly better than KNN.

The kappa value defines the fitness of the model as it shows the agreement between actual and predicted values and Landis and Koch states kappa value with more than 0.60 as a good classifier. According to Fleiss, kappa value greater than 0.75 is considered as an excellent value.

Table 1: Comparison of individual models

Model Name	Accuracy	Kappa Value
C5.0	72%	0.62
KNN	70.81%	0.60
Rpart	71%	0.62

In ensemble modelling, we are using simple averaging ensemble method which takes the average of all three predicted models. These three predicted models are KNN, C5.0 and Rpart. We have also calculated the accuracy of each individual model to evaluate the performance metrics of the individual models with comparison to ensemble model.

The ensemble modelling implemented with the help of averaging method uses all three base model's prediction along with the help of random forest algorithm which is used as a feature selection in ensemble modelling. Here, random forest is not implemented as the core base model but it contributes in bagging technique with the help of the ginni plot. The ginni plot represents the mean decreased accuracy value which helps in contributing the better model. The below diagram highlights the attributes which contributes in implementing ensemble model.

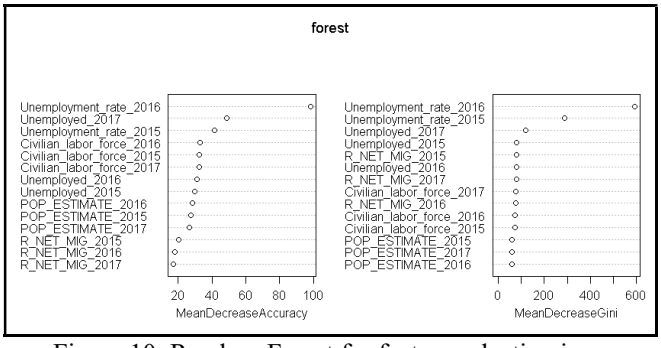

Figure 10: Random Forest for feature selection in ensemble modelling

The below diagram shows class imbalance with majority of the counties in class 2 with Unemployment rate in range of 3.5 to 4.5. which is a mediocre limit and relatively less counties lie in class 5 that has the highest unemployment range. Hence the Unemployment rate appears to be controlled and will decrease in coming years.

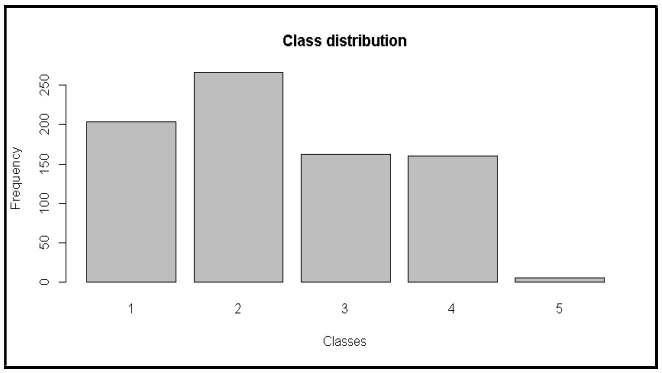

Figure 11: County distribution across the class

The below diagram shows the results of ensemble modelling with simple averaging method. The accuracy of the model is 83.81% with 0.78 kappa value which denotes that it outperforms all the individual models and gives better accuracy.

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1 182  23   1   0   0
         2   8 203  32   0   0
         3   0  21 151   2   0
         4   0   1  35 128   0
         5   0   0   0   6   4

Overall Statistics

               Accuracy : 0.8381
                 95% CI : (0.8107, 0.8631)
    No Information Rate : 0.3112
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7837
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class
Sensitivity            0.9579   0.8185   0.6895   0.9
Specificity            0.9605   0.9271   0.9602   0.9
```
Figure 12: Ensemble by Averaging method

In stacking technique, we are using gradient boosted method to improve the accuracy by eliminating the errors in the individual models. The accuracy given by GBM stacking model is 85.19% which is highest amongst all with the excellent kappa value of 0.80.

```
> confusionMatrix(test$Unemployment_rate_2017, gbm.pr
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1 188  18   0   0   0
         2  14 204  25   0   0
         3   0  29 138   7   0
         4   0   0  19 144   1
         5   0   0   0   5   5

Overall Statistics

               Accuracy : 0.8519
                 95% CI : (0.8254, 0.8759)
    No Information Rate : 0.3149
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8021
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class
Sensitivity            0.9307   0.8127   0.7582   0.9
Specificity            0.9697   0.9286   0.9415   0.9
```
Figure 13: Ensemble by stacking method

E. Evaluation

Once the data mining models are implemented we need to evaluate the results. The aim to forecast the population was done using ARIMA model with RMSE value of 0.00138 which shows that the model is fit and MAPE value of 0.0128. The forecasted values show the upwards trend which depicts the rise of population in US in 2018.

Further to arima results it was clear that the scale of population is increasing and the consequence of the increasing population on our national economical projection that is unemployment rate was observed using ensemble modelling which is an promising machine learning technique. The results of this model were obtained by implementing feature selection by random forest and implementation by C5.0,KNN and Rpart. The results were then observed by averaging method that gave 83.81% of accuracy with 0.78 kappa value. In addition to this, stacking technique was evaluated by using gradient boost method which gave 85.19% of accuracy with 0.80 of kappa value.

Later on the implications of population were observed on employment rate by using deep learning approach of RNN model where in the resultant output by means of keras library implementation predicted employment rate in resonance with the actual testing employment rate. As a result of which these paper caters population census forecast in time series domain on two primarily strong fields of national economic projections and the observed results are good as the RMSE value of the model was 0.37 which represents that the model performed well.

F. Knowledge

This is the last stage of KDD approach where the data insights were drawn from the pattern evaluated using the machine learning and deep learning models. The knowledge discovered was the rise in population of US in 2018 using ARIMA model. The RNN model analyzed the employment rate in US. Later on for analyzing the unemployment rate of US, we have used ensemble modelling.

IV. CONCLUSION & FUTURE SCOPE

The objective of this paper was to analyse the census dynamics based on major economic factors such as population, employment rate and unemployment rate. The aim of this paper was to forecast the US population in 2018 and predict the employment and unemployment rate of US using machine learning and deep learning techniques.

The models were evaluated based on the data used for modelling. We have used ARIMA model for population forecasting which predicts the rise in US population. The RNN model was used to analyse the employment rate of US were the model performed fairly decent and gave results that resonated with the desired test data. After that, we implemented classification model to analyse the unemployment rate of US in 2018. We used ensemble model with averaging and stacking technique to predict the unemployment rate. The ensemble stacking technique outperformed all the individual model performance with better accuracy and less errors.

The future work for this project can be done by tuning the RNN model in order to increase the performance of the prediction. The GRU model based on RNN can also be used which is the latest model used for time series model with less amount of data.

REFERENCES

[1] M. Popescu, "Modelling prediction of unemployment statistics using web technologies", *HOLISTICA – Journal of Business and Public Administration*, vol. 8, no. 3, pp. 55-60, 2017.

[2] S. Anderson, A. Cooper, O. Jensen, C. Minto, J. Thorson, J. Walsh, J. Afflerbach, M. Dickey-Collas, K. Kleisner, C. Longo, G. Osio, D. Ovando, I. Mosqueira, A. Rosenberg and E. Selig, "Improving estimates of population status and trend with superensemble models", *Fish and Fisheries*, vol. 18, no. 4, pp. 732-741, 2017.

[3] E. Morgenroth, "Evaluating Methods for Short to Medium Term County Population Forecasting," p. 30.

[4] D. Gawatre, M. Kandgule and S. Kharat, "Comparative Study of Population Forecasting Methods", *IOSR Journal of Mechanical and Civil Engineering*, vol. 13, no. 04, pp. 16-19, 2016.

[5] U. K. Mallick and M. H. A. Biswas, "Optimal control strategies applied to reduce the unemployed population," *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2017.

[6] W. Anggraeni, R. Vinarti and Y. Kurniawati, "Performance Comparisons between Arima and Arimax Method in Moslem Kids Clothes Demand Forecasting: Case Study", *Procedia Computer Science*, vol. 72, pp. 630-637, 2015.

[7] Y. Yang, C. Liu, and F. Guo, "Forecasting method of aero-material consumption rate based on seasonal ARIMA model," *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, 2017.

[8] Y. Zhang, J. Wang, Q. Zeng, H. Qiu, and H. Tan, "Near future prediction of European population through Chebyshev-activation WASD neuronet," in *2015 Sixth International Conference on Intelligent Control and Information Processing (ICICIP)*, 2015, pp. 134–139.

[9] K. Wangdi, P. Singhasivanon, T. Silawan, S. Lawpoolsri, N. White and J. Kaewkungwal, "Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: A case study in endemic districts of Bhutan", *Malaria Journal*, vol. 9, no. 1, p. 251, 2010.

[10] Tarno, Subanar, D. Rosadi, and Suhartono, "New procedure for determining order of subset autoregressive integrated moving average (ARIMA) based on over-fitting concept," *2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE)*, 2012.

[11] McNally. Sean, Roche. Jason and Caton. Simon, "Predicting the Price of Bitcoin Using Machine Learning", 339-343, 10.1109/PDP2018.2018.00060, 2018.

[12] Z. Che, S. Purushotham, K. Cho, D. Sontag and Y. Liu, "Recurrent Neural Networks for Multivariate Time Series with Missing Values", Scientific Reports, vol. 8, no. 1, 2018.

[13] E. Khosla, D. Ramesh, R. Sharma and S. Nyakotey, "RNNs-RT: Flood based Prediction of Human and Animal deaths in Bihar using Recurrent Neural Networks and Regression Techniques", Procedia Computer Science, vol. 132, pp. 486-497, 2018.

[14] "County-level Data Sets - Unemployment - Data.gov", Catalog.data.gov, 2018. [Online]. Available: https://catalog.data.gov/dataset/county-level-data-sets/resource/6117e794-f5f6-47b0-90d1-ab32272595b1?inner_span=TrueEmployment. [Accessed: 10- Jul- 2018].

[15] "County-level Data Sets - Population - Data.gov", Catalog.data.gov, 2018. [Online]. Available: https://catalog.data.gov/dataset/county-level-data-sets/resource/1d59ffe4-227e-4837-a22a-423593bbeed3. [Accessed: 11- Jul- 2018].

[16] "Working Age Population: Aged 15-64: All Persons for the United States", Fred.stlouisfed.org, 2018. [Online]. Available: https://fred.stlouisfed.org/series/LFWA64TTUSM647S. [Accessed: 11- Jul- 2018].

[17] "Employment Rate: Aged 25-54: All Persons for the United States", Fred.stlouisfed.org, 2018. [Online]. Available: https://fred.stlouisfed.org/series/LREM25TTUSM156S. [Accessed: 11- Jul- 2018].

[18] "Recurrent Layers - Keras Documentation", Keras.io, 2018. [Online]. Available: https://keras.io/layers/recurrent/. [Accessed: 23- Jul- 2018].

[19] Azevedo. Ana, Filipe. Santos and Manuel, "KDD, semma and CRISP-DM: A parallel overview", 182-185, 2008.

[20] M. Fayyad Usama, Piatetsky-Shapiro. Gregory and Smyth. Padhraic, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, 17. 37-54. 10.1609/aimag.v17i3.1230, 1996.

[21] R. Guruvayur, Sivaramakrishnan and R. Suchithra, "A detailed study on machine learning techniques for data mining", 1187-1192. 10.1109/ICOEI.2017.8300900, 2017.