WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Job Position Detection: A Data Science Approach

*Neha Kunjan Shah**

P.P. Savani University

ABSTRACT

Machine learning is one of the luxuriate areas of computer science, with momentous applications. It is the task of finding meaningful patterns in the data. Various research has been carried out for predicting average salaries of job positions. To detect Job, Position, based on average salary, given correlated explanatory variables that cover many aspects of Job activities on the internet. This research aimed at proposing a step-by-step process on how to come up with an approach of detecting Job Position, compares various learning algorithms, single classifiers, and ensemble classifiers are considered. Different machine learning algorithms were applied to the dataset and five were selected.

AdaboostClassifier(ADC),randomForestClassifier(RFC),GradientBoostingClassifier(GBC),GBoostClassifier(XGB),ExtraTreesClassifier (EXC), to form Heterogeneous Ensemble. To implement the algorithms, a dataset that contained 955 instances obtained from Glassdoor using web scraping was used for the classification. Furthermore, using correlated selected attributes as an independent variable and one as a dependent variable in the analysis. The results show that the Heterogeneous ensemble found to be the algorithm with the most precision and accuracy and higher Matthew's correlation coefficient. AdaBoostClassifier and Random Forest Classifier(RFC) algorithms were found to be the next most accurate after GradientBoostingClassifier(GBC) accordingly. Research shows that the time taken to build a model and precision is a factor, perhaps; while Heterogeneous Ensemble and Mean Square Error are other factors in different circumstances. Consequently, Machine Learning algorithms require training and testing accuracy, Precision, Recall, F1-Score, ROC-Score, Mean Square Error, and Root Mean Square Error, and lastly Matthew's Correlation Coefficient to possess supervised conjecturing machine learning.

Keywords: Data Science, Machine Learning, Heterogeneous Ensemble, Job Position, Multi-class Classification

1. Introduction

Ensemble modeling can be seen as a way to improve the accuracy of predictive or classification tasks. The ensemble can be seen as the cooperation of a group of models to solve the problem. The problem modeling process is very demanding because it involves many cooperating entities. One of the main challenges is that the state of the entity is constantly changing. You can change the state to adapt to new conditions, so you can be part of the collaboration or leave the collaboration. There are several methods that can be used for ensemble modeling, but it is important to combine them for better results. [1].

Jobs describes a person's journey through learning, work, and other aspects of life. The term career (work) has several meanings. It can be seen from various perspectives.. Job is the time spent on work and profession. In general terms, this can mean progress or upward movement towards a linear progression [2]. In machine learning, classification refers to the problem of predictive modelling that predicts the class name of a particular example of input data. Classification can be binary classification, multi-class classification, multi-label classification, and imbalance classification. Multi Classification was used for this task.

This paper, intends to finds the Heterogeneous Ensemble (Aggregating the detection results of model(classifiers) by summing the accuracy and taking the

Average mean of AdaboostClassifier (ADC), RandomForestClassifier (RFC),GradientBoostingClassifier (GBC), XGBoostClassifier (XGB), ExtraTreesClassifier, traTreesClassifier(EXC)).These learning algorithms (classifiers) are selected based on their performances learned from our research performed on the dataset obtained from Glassdoor using web scraping.

A. MACHINE LEARNING ALGORITHMS FOR CLASSIFICATION

a) Heterogeneous Ensemble:

In ensemble learning, we trained individual classifiers like decision trees or neural networks and then combine predictions of these individual classifiers for classifying new instances. Each member of the ensemble must complement the other. If the used ensemble methods are complementary, the probability to identify an error in the prediction increases as well as it is possible to correct this error with other methods. Instead of many traditional machine learning algorithms that generate a single model, ensemble learning methods generate multiple models. The ensemble gives each related model a new example, then receives those predictions and combines them accordingly. The ensemble classification performed in the present research is an aggregation of predictions of the multiple classifiers with the goal of improving accuracy[3].

Heterogeneous Ensembles (HEE) use different fine-tunes algorithms. It usually works well when the estimator is low. The number of algorithms should always be odd (3+) so that we can avoid ties. For example, we can combine SVM, a neural network with the voting system so that results will improve. Then use this improved classifier to classify new instances. There are different ways to combine models along with voting like taking the average, stacking to aggregate the results. [4].

b) AdaBoostClassifier (ADC):

In AdaBoost classification method, a weak learner will get called repeatedly in rounds. Poor learners are slightly better than random guessing and are learning algorithms that find the boundaries between two classes (positive and negative). In this, a strong learner will be created from number of weak learner so that we can achieve better separation in classes. The strong learner is a weighted majority vote of the weak learners[5].

The general idea behind the boosting method is to train the predictors one by one, each trying to modify its predecessor. AdaBoost and Gradient Boosting are two commonly used boosting algorithms. AdaBoost is similar to Random Forest in that it combines the predictions of individual decision trees in the forest to determine the final classification. AdaBoost is similar to Random Forest in that it combines the predictions of individual decision trees in the forest to determine the final classification. There are however, some subtle differences. For instance, In AdaBoost, the decision tree has a depth of 1 . In addition, the predictions made by each decision tree have different effects on the final prediction of the model. [6].

c) RandomForestClassifier (RFC):

Random forest is a classifier consisting of a collection of tree-structured classifiers in which independent random vectors are similarly distributed and each tree has a unified vote on the most popular one. at input x[7]. Random forests is a supervised learning algorithm. It is said that the more trees there are, the stronger the forest is. Random Forest creates decision trees based on randomly selected data samples, receives predictions from each tree, and votes for the best solution.. It's also a great indicator of the importance of functionality. The Random Forest has different uses like Recommendation engine, image classification, feature selection. This works in 4 steps.

Random sample selection from a specific dataset. Create a decision tree for each sample and get the prediction results from each decision tree. Vote for each predicted result.

Select the forecast result with the most votes as the final forecast [8].

d) GradientBoostingClassifier (GBC):

Gradient boosting is a method for developing classification and regression models to optimize the learning process of a model. The model learning process is mostly non-linear and more widespread. widely known as decision or regression trees[9].

A group of weak prediction models, for example, regression decision trees, are modelled by adding new learner in a gradual sequential manner. It consists of nodes and leaves that provide forecast results based on the decision node. Regression trees are individually weak models, but when viewed as an ensemble, they are significantly more accurate. Therefore, the ensembles are built gradually in an incremental manner such that every ensemble rectifies the error in the previous ensemble, mathematically as Eq.[10].

e) ExtraTreesClassifier (EXC):

ExtraTreesClassifier is an ensemble classifier used for feature selection. It is constructed using the training sample. It is a similar kind of Random Forest ensemble machine learning classifier.

Therefore, extra-treess Classifier implements a class of Meta estimators that fits number of mixed decisions tree on various reduced sample of the dataset and uses averaging to brush up the predictive accuracy and manage overflourish[11].

f) XGBoostClassifier(XBC):

XGBoost is an efficient and scalable machine learning classifier, which was popularized by Chen and Guestrin in 2016 [12]. The Gradient Boosting Decision Tree is the original model of XGBoost, which combines multiple decision trees in a boosting fashion. In general, each new tree is created with gradient reinforcement to reduce the rest of the previous model. The remainder is indicated by the difference between the actual value and the predicted value. The model is trained until the number of decision trees sets the threshold. XGBoost follows the same principles of gradient enhancement. Use the number of boosts, learning rate, subsampling ratio, and maximum tree depth to control overfitting and improve performance.More importantly, XGBoost optimizes the size of the function's target, tree size, and weights controlled by standard regulatory parameters. [13].

2. Related Work

Review of literature is essential step in the development of research project. It enables the researcher to develop perception into the research study and plan the methodology, Further, it provides the basis for future investigation, viability of the study, and specifies limitations of data collection. It helps to relate the detecting from one study to another study with a view to begin a comprehensive body of specific knowledge in a professional regulation, from which well-founded and appropriate theory may be developed [14].

Literature reviews can be a useful, critical, and useful summary of a particular topic. It helps identify known (and unknown) areas of the field, identify controversial or controversial areas, and formulate questions that require further investigation. [15].

Review of published and unpublished research and non-research literature is an integral part component of any scientific research. It involves a systematic identification, location, security and summary of written materials that contain information regarding a research problem. It broadens the understanding and gives an insight necessary for the development of a broad conceptual context into which the problem fits [16].

Conducted a study to determined Fake Job Recruitment Detection Using Machine Learning Approach, the dataset contained 17,876 instances obtained from the Kaggle website. The training and test dataset are taken from these data tuples independently and also the last attribute (classify) represents our result set (class label) which has values of "1" or "0". Every qualifying data value of the attributes of our dataset will result in "yes" and every inappropriate attribute value will result in "0". Different classifiers were used for checking fraudulent posts on the web and the results of those classifiers were compared for identifying the best employment scam detection model. Two major types of classifiers are examined, a single classifier and an ensemble classifier, to detect fraudulent classified ads. However, experimental results show that ensemble classifiers are the best classification for detecting fraud on individual classifiers [17].

Conducting a survey to determine the gradual contribution of complex problem-solving skills to work level, work complexity, and salary prediction). They investigated whether complex problem-solving skills (CPS) after managing GMA and education gradually contributed to professional success. The result shows that CPS offered no incremental increase in predicting job level. CPS appears to be related to work complexity and salary in many professions, and this relationship cannot be explained as a product of GMA and education. As a result, CPS can step-by-step predict success, contribute to the theory of gravity in the workplace, and contribute to understanding the complex cognition of the workplace. [18].

Performed a study title Salary Prediction in the IT Job Market with Few High-Dimensional Samples. The research contained 4,000 instances of the dataset collected from the Spanish IT recruitment portal, that help in identify the most rewarded and demanded items in job offers which is the key for recruiters and candidates. This work summarized that experience is more rewarded than education, identified five profile clusters based on required skills, and lastly developed an accurate salary-range classifier by using tree-based ensembles [19].

Submitted Papers To use cognitive proficiency tests to predict long-term work performance, we collect data from international technology companies with more than 3,000 employees, and the results of proficiency tests are objective and proficient. Evaluated how it relates to both subjective work performance measurements. Some factors such as Supervisory performance ratings, level of promotion, and salary increase significantly attributed to variance in test scores; nevertheless, these results were inconsistent. Most training courses did not have a remarkable relationship with test scores. In addition, the nature of sales did not ease the relationship between proficiency test results and job performance. Studies show that aptitude test results are related to long-term job performance factors, but other factors make up the majority of the variance and are considerations when predicting long-term job performance. Shows that compatibility is not the only thing. [20].

Performed research predicting the salary satisfaction of exempt employees, the study observed the degree to which salary satisfaction can be predicted using company collected and maintained information commonly available to salary administrators. Different predictors are examined which include years of continuous services, educational level, annual performance rating, an estimate of career potential, monthly salary, a measure of the most recent salary increase, and employee gender. Different hypotheses derived from Lawler's (1971) model of pay satisfaction also were conducted, focusing on the relative contribution of perceived performance, perceived job demands, certain non - monetary outcomes, and external and internal pay equity. The research minimized a sample of managerial, professional, and technical employees from a large national oil company. Observations suggest that salary and gender are key objective predictors, as only a small portion of wage satisfaction can be taken into account, without including broad employee perceptions. Perceived performance, perceptions regarding supervision, advancement opportunity, and the company's benefits package, and both external and internal pay equity, were attributed to paying satisfaction in the direction predicted by Lawler's model [21].

All figures should be numbered with Arabic numerals (1,2,3,….). Every figure should have a caption. All photographs, schemas, graphs and diagrams are to be referred to as figures. Line drawings should be good quality scans or true electronic output. Low-quality scans are not acceptable. Figures must be embedded into the text and not supplied separately. In MS word input the figures must be properly coded. Lettering and symbols should be clearly defined either in the caption or in a legend provided as part of the figure. Figures should be placed at the top or bottom of a page wherever possible, as close as possible to the first reference to them in the paper.

The figure number and caption should be typed below the illustration in 8 pt and left justified [***Note:*** one-line captions of length less than column width (or full typesetting width or oblong) centered]. For more guidelines and information to help you submit high quality artwork please visit:http://www.elsevier.com/wps/find/authorsview.authors/ authorartworkinstructions. Artwork has no text along the side of it in the main body of the text. However, if two images fit next to each other, these may be placed next to each other to save space. For example, see Fig. 1.

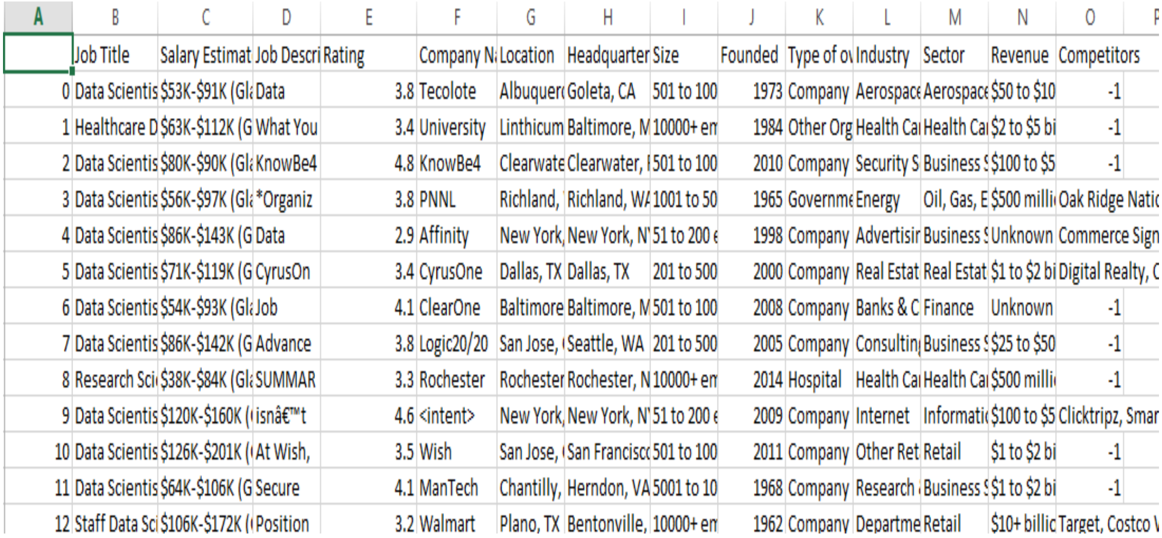
3. Methods

This phase provided a step-by-step explanation of how the project was executed. This work is implemented in six phases of CRISPDM. CRISPDM represents the cross-industry process of data mining. The CRISPDM methodology provides a structured approach for planning data mining projects. This is a robust and proven method. We do not claim ownership of it. I could not do it. However, it does advocate powerful utilities, flexibility, and usefulness in solving sensitive business problems using analytics. This is a common thread that runs through almost all customer relationships. This model is an idealized set of events. In reality, many of the tasks can be performed in a different order, so you often need to go back to the previous task and repeat certain actions. [22]. Purpose of the research.

Much research has been done to predict the average salary of a job. This study aims to classify these jobs based on average salary, in addition to the correlation characteristics available in the dataset. Looking at the details of the characteristics of well-known work-related datasets edited by investors, the detection and classification of job positions as a multi-class classification is done with the help of data analysis and machine learning algorithms. Positions are categorized as "Data Scientist", "Data Engineer", "Analyst", "Machine Learning", "Manager", "Director", and finally all other positions are categorized as "Other Positions". Therefore, you may be able to use the same algorithm and model. Can be modified for larger datasets.

a) Data Understanding

The second stage of the CRISPDM process needs to collect the data listed for this task. This initial collection involves loading the data when needed to understand it. This includes the process of collecting datasets by web scraping, manual collection, or downloads from specific websites. In this article, we will use web scraping (Selenium Web Scraper) to retrieve records from the Glassdoor website



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	Job Title	Salary Estimat	Job Descri	Rating	Company Na	Location	Headquarter	Size	Founded	Type of ov	Industry	Sector	Revenue	Competitors	
0	Data Scientis	$53K-$91K (Gla	Data	3.8	Tecolote	Albuquer	Goleta, CA	501 to 100	1973	Company	Aerospace	Aerospace	$50 to $10	-1	
1	Healthcare D	$63K-$112K (G	What You	3.4	University	Linthicum	Baltimore, M	10000+ em	1984	Other Org	Health Ca	Health Ca	$2 to $5 bi	-1	
2	Data Scientis	$80K-$90K (Gla	KnowBe4	4.8	KnowBe4	Clearwate	Clearwater,	501 to 100	2010	Company	Security S	Business S	$100 to $5	-1	
3	Data Scientis	$56K-$97K (Gla	*Organiz	3.8	PNNL	Richland,	Richland, WA	1001 to 50	1965	Governme	Energy	Oil, Gas, E	$500 milli	Oak Ridge Natio	
4	Data Scientis	$86K-$143K (G	Data	2.9	Affinity	New York,	New York, N	51 to 200 e	1998	Company	Advertisir	Business S	Unknown	Commerce Sign	
5	Data Scientis	$71K-$119K (G	CyrusOn	3.4	CyrusOne	Dallas, TX	Dallas, TX	201 to 500	2000	Company	Real Estat	Real Estat	$1 to $2 bi	Digital Realty, C	
6	Data Scientis	$54K-$93K (Gla	Job	4.1	ClearOne	Baltimore	Baltimore, M	501 to 100	2008	Company	Banks & C	Finance	Unknown	-1	
7	Data Scientis	$86K-$142K (G	Advance	3.8	Logic20/20	San Jose,	Seattle, WA	201 to 500	2005	Company	Consultin	Business S	$25 to $50	-1	
8	Research Sci	$38K-$84K (Gla	SUMMAR	3.3	Rochester	Rochester	Rochester, N	10000+ em	2014	Hospital	Health Ca	Health Ca	$500 milli	-1	
9	Data Scientis	$120K-$160K (	isnâ€™t	4.6	<intent>	New York,	New York, N	51 to 200 e	2009	Company	Internet	Informati	$100 to $5	Clicktripz, Smar	
10	Data Scientis	$126K-$201K (	At Wish,	3.5	Wish	San Jose,	San Francisc	501 to 100	2011	Company	Other Ret	Retail	$1 to $2 bi	-1	
11	Data Scientis	$64K-$106K (G	Secure	4.1	ManTech	Chantilly,	Herndon, VA	5001 to 10	1968	Company	Research	Business S	$1 to $2 bi	-1	
12	Staff Data Sci	$106K-$172K (	Position	3.2	Walmart	Plano, TX	Bentonville,	10000+ em	1962	Company	Departme	Retail	$10+ billic	Target, Costco V	

Fig. 1. Schema Structure of the dataset

From the above table, there are six (6) columns, selected to carry out the work using correlations between dependent and independent variables with their meaning and purposes.

a) Independent Variables/Explanatory Variables

1.Salary Estimate: It is a float value that defines the salary ranges for a particular job title called “average salary”

2.Size: integer storing the number of staff available in the company.

3.Revenue: Float defines the capacity of turned out that a company is able to gain in a year.

4.Job Description: The feature is integer storing the length of the character of work writing to an employee.

5.Rating: Float which stores the sentiment value of the people interested in the company.

6.Company Name: Categorical variable feature which stores the names of the companies.

7.State: categorical variable storing the

8. locations of the companies present.

9.Dependent/Response variable

10. Label/Target: Categorical string that stores the name Jobs title members.

There is a total of 955 instances where 358 “data scientist”, 238 “other Job positions”, 158 “data engineer”, 124 “analyst”, 36 “manager”, 26 “Machine Learning” and 16 “directors” categorized and used for classifications from “Job title” attribute in the dataset.

c) Data Preparation

In machine learning, data preparation can also be defined as data preprocessing. Data preprocessing allows data cleansing to remove unwanted data. This allows users to use a dataset that contains more valuable information for later manipulation of the data in the data mining process after the preprocessing phase. In short, data preparation is the process of transforming raw data so that data scientists and analysts can run machine learning algorithms to generate insights and make predictions.

This task removes the delimiter from some attributes, gets the lower and upper bounds, and finds the average of those values. Use the interquartile range to remove outliers by typing the missing "last known value" back and forth and removing observations that contain three or more outliers. Convert continuous attributes to bins to significantly improve performance while reducing class imbalances, remove columns whose correlation matrix does not relate to the target label, and finally use the label encoding process to categorize labels. To a number.

d) Modelling and Evaluation

Model evaluation is an integral part of the model development process. This will help you find the best model to represent your data and how well the selected model will work in the future. Evaluating the performance of a model using the data used for training is unacceptable in data science because it can easily generate an overly optimistic and overfitting model.

This paper divides the data, creates an ensemble of the same type as a separate model for each machine learning algorithm, improves model performance, and creates heterogeneous ensembles using the most powerful machine learning algorithms. increase. I have created a training set and a test set to check the distribution. Analysis of results by mutual verification by kfold. We used the

data split holdout method to create training and testing departments suitable for the classification model. All models were created with two split dates to check variance and performance. Two data splits are created twice, one using the data from AdaboostClassifier (ADC), RandomForestClassifier (RFC), GradientBoostingClassifier (GBC), XGBoostClassifier (XGB), ExtraTreesClassifier (EXC) to create a heterogeneous ensemble (conversion). And normalization)

Because this is a multi-class classification issue. Evaluate the performance of each model using the confusing matrix of the following metrics.

Accuracy: Ratio of true results to total number of cases investigated TP + TN / (TP + TN + FP + FN).

Accuracy: The expected percentage of positives is actually positive TP / (TP + FP).

Recall / Sensitive: The actual positive percentage is correctly classified as TP / (TP + FN).

Specificity / 1 False Positive Rate: The actual negative rate is correctly classified as TN / (TN + FP).

F1 Score: Average precision and recall. Find the balance between two metrics that are very different from each other (2 * fit rate * recall rate) / (match rate + recall rate).

ROC: The AUC ROC curve is a performance indicator for classification problems at various threshold settings. ROC is a probability curve and AUC is a measure or measure of repairability. This shows how well the model can distinguish

classes. The higher the AUC, the more the model can predict 0 as 0 and 1 as 1. By analogy, the following applies: The higher the AUC, the better the model can distinguish between patients with and without disease. (Wikipedia)

MSE: In statistics, the mean square error (MSE) of an estimator (a method of estimating an unobserved amount) measures the mean square of the error-that is, the root means square difference between the estimates. What is presumed to be. MSE is a risk function that corresponds to the expected value of squared error loss. The fact that MSEs are strictly positive (non-zero) in most cases is due to randomness, or because the estimator does not take into account information that can provide more accurate estimates. (Wikipedia)

RMSE: Root Mean Square Deviation (RMSD) or Root Mean Square Error (RMSE) is the value predicted by the model or estimator (sample value or population value) and the observed value (Wikipedia).

Matthews Correlation Coefficient: This is the ratio of the properties of binary and multi-class classifications. It takes into account the positive and negative consequences of correctness and is generally considered a balanced measure that can be applied to very different class sizes.

MCC is basically a correlation coefficient value between 1 and +1. A factor of +1 is a complete prediction, 0 is an average random prediction, and 1 is a reverse prediction. Statistics are also known as the Phi coefficient. [Source: Wikipedia].

4. Results

In this section, we are going to put down the results and analysis of the whole research conducted, which shows the tables and figures of the analysis and results of the nineteen (19) classification algorithms (Classifiers) applied in our dataset.After Cross Validation is performed to form a heterogeneous ensemble, an algorithm (classifier) that performs better in terms of test accuracy is selected. Below is a screenshot of the data analysis performed on the dataset in relation to the performance metrics.

MLA Name	**Train Accuracy**	**Test Accuracy**	**MLA Time**	**Precision**	**Recall**	**F1-Score**	**ROC Score**	**MSE**	**RMSE**	**MCC**
MLA Name	**Train Accuracy**	**Test Accuracy**	**MLA Time**	**Precision**	**Recall**	**F1-Score**	**ROC Score**	**MSE**	**RMSE**	**MCC**
XGB classifier	1.000	1.0000	0.1331	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	1.0000
Ridge Classifier CV	0.6204	0.6354	0.0024	0.2223	0.1833	0.2857	0.6036	0.7291	0.8539	0.5134
Random Forest classifier	1.0000	0.9896	0.1554	0.9726	0.9960	0.9535	0.9756	0.0520	0.2282	0.9861
Quadratic Discriminant Analysis	0.2448	0.2656	0.0013	0.0599	0.0379	0.1429	0.5000	6.1667	0.2483	0.0000
Perceptron	0.2775	0.3021	0.0059	0.1399	0.1679	0.2362	0.5583	3.2919	1.9804	0.0000
Logistic Regression CV	0.8665	0.8958	2.6632	0.6435	0.6340	0.6611	0.8206	0.2344	0.4841	0.8609
LIinearSVC	0.6636	0.6927	0.1650	0.3316	0.3524	0.3512	0.6409	0.8594	0.9270	0.6149
Lineardiscriminant Analysis	0.4568	0.4531	0.0015	0.2024	0.1960	0.2247	0.5544	3.4115	1.1847	0.2165
K Neighbors Classifier	0.6571	0.5000	0.0008	0.2923	0.2989	0.2928	0.5974	2.9219	1.7094	0.3156
Gradient Boosting Classifier	1.0000	1.0000	0.5223	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
GaussianNB	1.0000	1.0000	0.0009	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	1.0000
Extra Tree Classifier	1.0000	0.8299	0.0006	0.7504	0.8075	0.7538	0.8606	1.1094	1.0533	0.7642
Decision Tree Classifier	1.0000	1.0000	0.0007	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	1.0000
Bagging Classifier	1.0000	1.0000	1.0000	0.0168	1.0000	1.0000	1.0000	0.0000	0.0000	1.0000
AdaBoost Classifier	0.8220	0.8443	0.0842	0.5267	0.499	0.5714	0.7715	0.1719	0.4146	0.8052

A) Cross Validate Of Models

We compared thirteen (13) most popular classifiers and evaluate the test accuracy of each of them by a stratified k-fold cross-validation procedure, ten (10) K-fold.

After Cross-Validation with ten (10) K-fold, we choose algorithms that perform best in test accuracy, five algorithms were selected from the above table. These are AdaBoost, ExtraTrees, Random Forest, Gradient Boosting, and XGBoost Classifier, which are used for performing an Ensemble method. The figures below show models plotted for learning performance, experience over time. However, the classifiers reveal the best of their learning curve. The five classifiers give plus or minus the same prediction, but there are some differences. These differences between the five classifiers predictions are sufficient to consider an Ensemble vote. We tendered to pass the argument "soft" to the voting parameter to take into account the probability of each vote.

B) Actual and Predicted

The model was successfully developed, the voting classifier was examined that gave more test and train accuracy, Precision, Recall, and Matthew's Correlations Coefficient. The results of the first ten (10) instances were generated below:

TABLE III. ACTUAL AND PREDICTED RESULT

S/N	Actual	Predicted
0	3	3
1	2	2
2	2	2
3	0	0
4	0	0

5	4	4
6	0	0
7	3	3
8	3	3
9	3	3

References

1. A. Petrakova, M. Affenzeller, and G. Merkurjeva, “Heterogeneous versus Homogeneous Machine Learning Ensembles,” Inf. Technol. Manag. Sci., vol. 18, no. 1, 2016, doi: 10.1515/itms-2015-0021.

2. “Career: Definition, Career Patterns, Career vs Job.” https://www.iedunote.com/career (accessed Jan. 05, 2021).B. Smith, “An approach to graphs of linear forms (Unpublished work style),” unpublished.

3. Z.-H. Zhou, Ensemble Methods, Foundations and Algorithms.pdfD

4. Giorgos Myrianthous, “machine learning - Homogeneous vs heterogeneous ensembles,” 2020. https://stackoverflow.com/questions/49445446/homogeneous-vs-heterogeneous-ensembles (accessed Dec. 30, 2020).

5. H. Fleyeh and E. Davami, “Multiclass Adaboost Based on an Ensemble of Binary AdaBoosts,” Am. J. Intell. Syst., vol. 3, no. 2, pp. 57–70, 2013, doi: 10.5923/j.ajis.20130302.02.

6. 2019 Maklin, Cory, “AdaBoost Classifier Example In Python | by Cory Maklin | Towards Data Science.” https://towardsdatascience.com/machine-learning-part-17-boosting-algorithms-adaboost-in-python-d00faac6c464 (accessed Dec. 31, 2020).

7. E. Goel and E. Abhilasha, “Random Forest: A Review,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 7, no. 1, pp. 251–257, 2017, doi: 10.23956/ijarcsse/v7i1/01113.

8. A. Navlani, “Random Forests Classifiers in Python - DataCamp,” Datacamp, 2018. https://www.datacamp.com/community/tutorials/random-forests-classifier-python?tap_a=5644-dce66f&tap_s=951023-a33697&utm_medium=affiliate&utm_source=kaybajaj&tm_subid1=213232&tm_subid2=20200708cleeprchv4cx.

9. N. Chakrabarty, T. Kundu, S. Dandapat, A. Sarkar, and D. K. Kole, Flight arrival delay prediction using gradient boosting classifier, vol. 813, no. September 2018. Springer Singapore, 2019.

10. “View of SMOTE_ Synthetic Minority Over-sampling Technique.pdf.” .

11. P. Harshalatha and R. Mohanasundaram, “A New Hybrid Strategy for Malware Detection Classification with Multiple Feature Selection Methods and Ensemble Learning Methods,” Int. J. Eng. Adv. Technol., vol. 9, no. 2, pp. 4013–4018, 2019, doi: 10.35940/ijeat.b4666.129219.

12. A. K. Agarwal, S. Wadhwa, and S. Chandra, “Diagnosis of tuberculosis--newer tests.,” J. Assoc. Physicians India, vol. 42, no. 8, p. 665, 1994.

13. K. Davagdorj, V. H. Pham, N. Theera-Umpon, and K. H. Ryu, “Xgboost-based framework for smoking-induced noncommunicable disease prediction,” Int. J. Environ. Res. Public Health, vol. 17, no. 18, pp. 1–22, 2020, doi: 10.3390/ijerph17186513.

14. F. G. Abdellah and E. Levine, “Better patient care through nursing research,” Int. J. Nurs. Stud., 1965, doi: 10.1016/0020-7489(65)90013-1.

15. A. Bolderston, “Writing an Effective Literature Review,” J. Med. Imaging Radiat. Sci., vol. 39, no. 2, pp. 86–92, 2008, doi: 10.1016/j.jmir.2008.04.009.

16. C. Li and D. Stacks, “Chapter 3. Research Methodology,” Meas. Impact Soc. Media Bus. Profit Success, pp. 32–42, 2016, doi: 10.3726/978-1-4539-1590-5/12.

17. S. Dutta and S. K. Bandyopadhyay, “Fake job recruitment detection using machine learning approach,” SSRG Int. J. Eng. Trends Technol., vol. 68, no. 4, pp. 48–53, 2020, doi: 10.14445/22315381/IJETT-V68I4P209S.

18. J. Mainert, C. Niepel, K. R. Murphy, and S. Greiff, “The Incremental Contribution of Complex Problem-Solving

19. Skills to the Prediction of Job Level, Job Complexity, and Salary,” J. Bus. Psychol., 2019, doi: 10.1007/s10869-018-9561-x.

20. I. Martín, A. Mariello, R. Battiti, and J. A. Hernández, “Salary prediction in the IT job market with few high-dimensional samples: A Spanish case study,” Int. J. Comput. Intell. Syst., vol. 11, no. 1, pp. 1192–1209, 2018, doi: 10.2991/ijcis.11.1.90.

21. S. G. Alexander, “Predicting long term job performance using a cognitive ability test,” ProQuest Diss. Theses, 2007.