**Project 4 – Machine Learning**

You are hired by one of the leading news channels CNBE who wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

Dataset for Problem: [Election\_Data.xlsx](https://olympus.greatlearning.in/courses/23976/files/2501827/download?verifier=WvQmxTtKou0as8KJzuOquae8eHYASkKNiISvpO7a&wrap=1)

* You have to submit 2 files :   
  1. **Answer Report**: In this, you need to submit all the answers to all the questions in a sequential manner. **It should include a detailed explanation of the approach used, insights, inferences, all outputs of codes like graphs, tables, etc.** Your report should **not** be filled with codes. You will be evaluated based on the business report.
  2. **Jupyter Notebook file/ R codes**: This is a must and will be used for reference while evaluating
* Please ensure timely submission as a post-deadline assignment will not be accepted.

**Data Ingestion:**

1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.   
1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

**Data Preparation:**

1.3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30).

**Modeling:**

1.4 Apply Logistic Regression and LDA (linear discriminant analysis).  
1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results.  
1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting.   
1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.

**Inference:**

1.8 Based on these predictions, what are the insights? (5 marks)

| **Criteria** |
| --- |
| **1.1) Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head() .info(), Data Types, etc . Null value check, Summary stats, Skewness must be discussed.** |
| **1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.** |
| **1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not?( 2 pts), Data Split: Split the data into train and test (70:30) (2 pts). It is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling. Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get\_dummies(drop\_first=True)) Data split, ratio defined for the split, train-test split should be discussed.** |
| **1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both model s (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)** |
| **1.5) Apply KNN Model and Naïve Bayes Model . Interpret the inferences of each model (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)** |
| **1.6) Model Tuning , Bagging (and Boosting . Apply grid search on each model (include all models) and make models on best\_params. Define a logic behind choosing particular values for different hyper-parameters for grid search. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.** |
| **1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model, classification report. Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model** |
| **1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, You should not give any technical suggestions. Evaluation will be considered if the recommendations are correct and business specific.** |
| **Quality of Business Report** |

**FAQs**

**Problem 1**

* **The data set is in xslx format so is it necessary to convert it into CSV?**

*when you load data to notebook only difference would be read\_csv or read\_excel depending on the file type you are reading. Post this read, it is a data frame to work upon and it doesn't matter which type of file data is read.*

* **In the first problem, the participants have assessed several parameters on a scale of 1 to 5. Here, is that 5 stands for a high score (for e.g., a participant who gives a score of 5 for the current national economic translates to him/her perceiving the economic conditions to be very good). Is this true, or is it the other way round? Kindly clarify?**

*Yes, it is right. 1 represents low and 5 represents a high score.*

* **What exactly do we need to consider for model tuning? Do we need to do tuning for all the models?**

*you can use GridSearchCV with different values of hyperparameters for tuning and overfitting issues.*

* **The initial check reveals that there are 8 duplicate records. Should we really drop them, because the question says that they surveyed 1525 persons and there are only 1525 records? Is there are a possibility that this is a mere coincidence of people having similar attributes (including age) rather than actual duplicates?**

*These duplicates need to be dropped because they do not add any value to the study, be it associated with different people.*

* **What is expected from problem1- 1.4 and 1.5? How it is different from 1.7?**

*Q 1.4, 1.5:- you need to make the models and write the basic information like recall, accuracy, etc. Comment for overfitting and underfitting issues. Explain the hyper-parameters you used in your basic model and the reason behind the selected values or default values.*

*Q1.7:- You need to write a detailed explanation like AUC/ROC, classification report, confusion matrics, model performance parameters like recall, accuracy, precision for both basic and tuned models using GridSearchCV.*

*Model Tuning :*

*This applies to all models including bagging and boosting models. For Example*

*1. KNN. Tuning is to choose the best k value, distance, and weights*

*2. Naive Bayes: Prior Probabilities*

*3. Logistic Regression: Different Solvers, C values, and regularization parameters*

*4. LDA: Prior Probabilities, Different algorithms*

* **What exactly is expected from model tuning?. Has SMOTE considered a technique of model tuning? Is every model supposed to be tuned? Random Forest itself uses bootstrap bagging, thus it itself represents bagging. Is it correct?**

*GridsearchCV is what we expect as model tuning. And every model needs to be tuned. You need to use RF and Bagging as separate models. You can use SMOTE if you think there is a class imbalance problem and prepare a different model and compare it with the original one.*

* **What are we expected to do in "Model Tuning" ? Just explain the concept? As we are any which ways checking on the performance of different models to suggest the best.**

*you are expected to explain the steps that you adopt for model tuning and the reasons behind choosing those parameters to tune the models. You may also use hyperparameters only for certain models, so in that case, the reason behind this would also be expected.*

*Model tuning means finding better parameters for the model and not just use the default values. For what parameters can be changed refer to the algorithm documentation.*

*Each ML algorithm has its own parameters to tune the model. Please check the grid search implementation in mentoring session notebooks. Ex: In random forest, two important hyperparameters are n\_estimators,max\_features*

* **The independent variables apart from age and gender are all ordered. If I do label encoding, then the original order gets changed. In this scenario should encoding be done? if we do one hot encoding then it will result in more than 30 columns as each, is that ok, since the dataset is small? Please advice.**

*it would be better to do one-hot encoding for the independent variables gender and age rather than label encoding. While doing the dummies section, you can explicitly mention which variable you want to create a dummy. by doing this, it will affect the other variables.*

* **Should we consider converting the ratings on economic condition national, household, Hague, Blair into categorical types?**

*It would be better to have the variables in ordinal data types as most of the machine learning models would take categorical independent variables into consideration.*

* **If both the Classes are equally important, then Which Class can be given label 1 and which one is 0? Which Performance Metrics are critical for model evaluation?**

*The class to be given as label 1 is based on the business problem that you want to prove/predict the model for. Both Recall and Precision can be equally important in this case.*

*Selecting the performance metrics depends on whether the model is part of supervised learning or unsupervised learning. Accordingly, the performance models can be chosen based on all methods that you have learned so far.*

* **The features such as 'economic.cond.national', 'economic.cond.household', 'Blair', 'Hague', 'Europe', 'political.knowledge' has dtypes is int64. Is it necessary to convert int64 into an object or we can proceed with int64?**

The machine will pull and load data and define the data type that is most suitable for the values in that column. It is our job to make sure that the correct data type is set for processing, especially when it comes to the point when categorical values having numbers are imported as integers by default. So yes it would be logical to check the data type by looking at the unique values of the column and see if they should actually be numeric or should be treated and worked as categorical.

* **Problem has ordinal values ranging from 1-11,1-5, and even 0-3. Should we encode them or treat them as continuous values?**

*Treat the values as it is and build the model for prediction, no need for a change.*

* **Problem has features that are ordinal but the rating scales are different. Since Logistic regression is sensitive to scaling, is it necessary to normalize all such features to bring to a similar rating scale (or) do we need to encode such features before various models are applied?**

*It would be better to normalize the rating scale before applying Logistic regression. When it comes to LDA, a suggestion would be to do the encoding.*

* **If by using GridSearchCV we get poorer performance values for recall, F1-score, etc from the tuned model than the base model which has no hyperparameters, should we go ahead and choose the base model as our final model? But In GridSearchCV we also use the cross-validation to get the best optimal model, so can we interpret this in the following way: Although the performance is poorer than the base model, it is still an optimized model and will perform better on the unseen data since cross-validation was performed on it along with probable hyperparameter values. So we should go ahead with the model formed using gridsearchCV although its performance is slightly less than the base model. Kindly tell me if this is correct?**

*Cross-validation can also be performed on the base model. See here we want to see your call while selecting a model, if you think the base model is good go for it.*

* **For questions 1.4, 1.5, & 1.6 models are built in jupyter notebooks, and what needs to be shown in the project report?**

*All workings will be done as part of the jupyter notebook, in the project report, you can explain the steps involved in building the model, attributes(Hyper-parameters) that you selected, and the reason behind it. Incase of tuning parameters being used, you can explain why you have thought to use it and how it has helped in improving the model. Comment on the values that you took for these parameters, overfitting-underfitting, etc.*

* **Do we need to create data points in case of an under-sampling? Not mentioned in the steps so not sure whether the same has to be done.**

*If you see the dataset is imbalanced and thus want to do a better tuning check based on balanced data then you should try SMOTE and see the results on the SMOTE dataset. Make sure you apply SMOTE only to train data and not to test data. Then you compare the test results between using the original data Vs SMOTE train data and see which way the model is more generalized.*

* **Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. for this, we have to use the random forest as a base estimator for bagging. For boosting we use the ADA boost and gradient boost should we use any base estimator or without any base estimator.**

*You are correct, base estimator would be required for bagging, bagging classifier should also be used here along with a random forest. For boosting you can use ADA and gradient boosting without a specific base estimator.*

* **In problem set, we have few outliers in only two features and the feature set is ordinal, so should we treat them.**

*Yes, outliers are to be treated for all analyses.*

* **Rather than doing a complete EDA code separately, can we do a panda's profile and infer the results from there in the submission report?**

*Pandas profile report is one simple to get all EDA done directly. But as part of the assignment requirement, you need to give each line of code and analysis separately. So for your own analysis, you can use pandas profile report, but for assignment submission, detailed EDA has to be done.*

* **Can we scale only one feature in the data set and leave remaining as it is, as they are already encoded kind of values? Is this the correct approach or not, please do a guide on this.**

*Scaling as an option is done if you have continuous nature of variables in the data set with different measures, but when you want to model, data with different measures would give us incorrect answers. So in case, the data points for independent variables are binary or ordinal in nature already, you can skip the step of scaling, else the process has to be adopted.*

* **Should I convert the target variable to (0,1) as this is a classification problem and I'm getting the same results for labels Labour and Conservative without converting?**

*There won't be any difference in encoding your target variable. You can keep it as it is or you can encode the target variable.*

* **If and when we do Scaling, we surely scale continuous variables like Salary, age, and so on 2. But what about coded Categorical variables. Especially if one variable has 2 categories and some other has 15. Should we scale these too?**

*Scaling is only for the continuous variables. Scaling on categorical variables does not make any sense.*

* ***So, only 'Age' variable should be scaled? Or How about transforming it into categories of Age bracket?***

*You are most welcome to try binning. Please try and see if any changes to model performance.*

* **All the variables in the election dataset are categorical except age. Though they are categorical, they are varying in scales. As LDA assumes that each input variable has the same variance, do we have to scale the variables?**

*LDA needs the data with the same variance but it is very difficult to make categorical data as normally distributed data. So please do not scale the categorical columns and build LDA.*

* **Can you explain more about the 'Blair' and 'Hague' variables of the Project? I am unable to understand what "Assessment of a leader" means here. Are voters evaluating the leaders here 1 being the least evaluation score n 5 the highest? Same for economic conditions?**

*Blair is Tony Blair (Name of the person Contesting from Labour Party) and Hague is the name of the person from the conservative party. Yes. Assessment scores. Lower 1 and higher is 5.*

* **Regarding model tuning for the KNN model, I am confused about which value to be used as an input parameter for weights. as if we use uniform, the results are entirely different from if we distance as weights. KNN is a distance-based algorithm, so should we consider distance as the correct option to be used as an input parameter. pls, let us know?**

*Weights can take two values. Uniform and distance. The Distance input for weights is different from a metric distance (Euclidian , Manhattan). Ex : You had given k value as 5 For nearest neighbor calculation, the algorithm uses the metric distance, After identifying the 5 nearest neighbors , do you want to give the equal weights for all the neighbors or any distance based weightage for generating the predictions.*

* **In Problem: Elections, the question is like “ Model Tuning, Bagging (Random Forest should be applied for Bagging) and Boosting.”Model tuning means it is using SMOTE or only hyperparameters, please be specific.**

*SMOTE is not a tuning algorithm. SMOTE is used to treat class imbalance. If your dataset is a class imbalanced dataset and it is affecting your performance metrics, please use SMOTE Tuning parameters is model related and will be used to tune the performance metrics. Ex : In random forest, the model tuning parameters are n\_estimators and max\_features.*