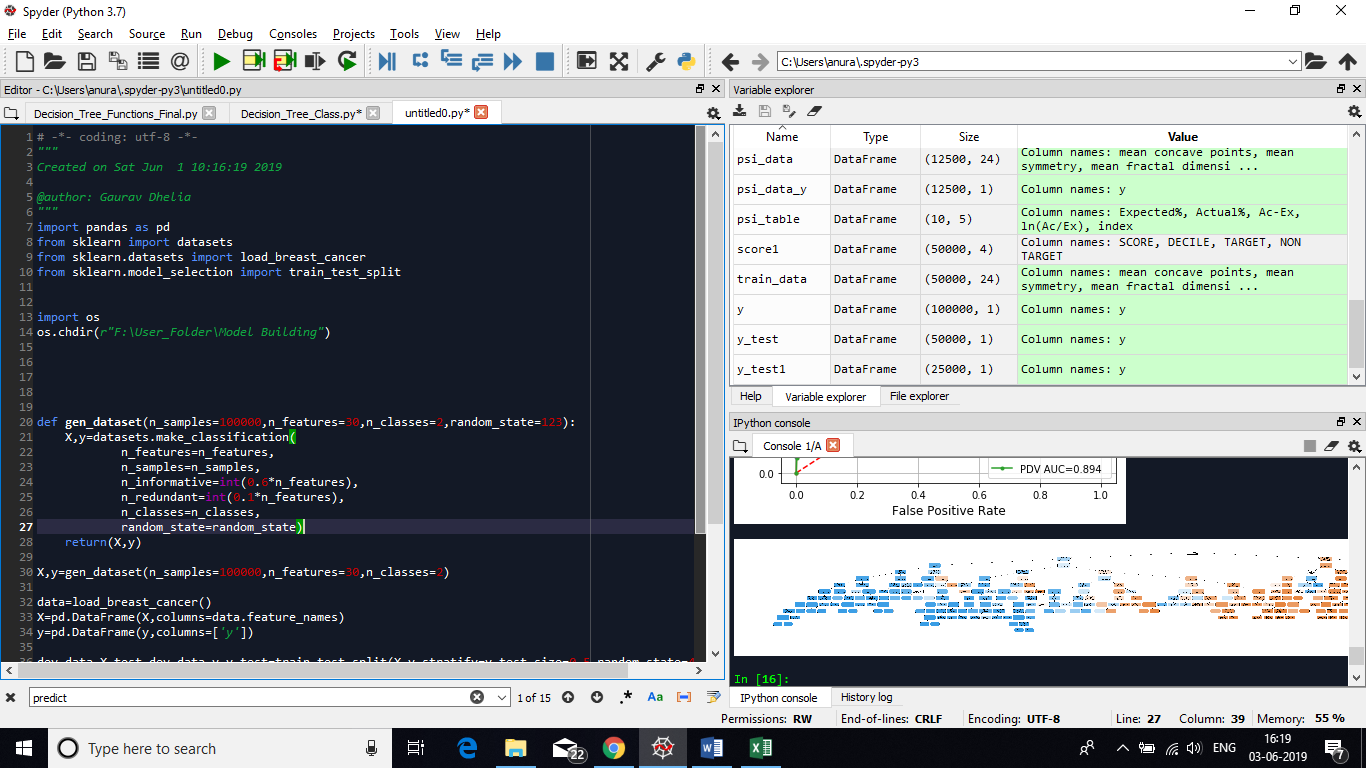
Functionalities of the Random Forest Classifier module:-

* Random Forest is an ensemble learning method for classification that operates by creating a multitude of decision trees at training time and outputting the class that is the mode of the classes. Random Forest corrects the Decision Trees’ habit of overfitting. It is a model wherein a group of weak models combine to form a powerful model.
* It splits the development data into train and test using stratified sampling based on the user’s choice and the fraction specified by the user.
* It takes care of categorical columns i.e. create dummy variables for each category.
* It fills the null values of continuous variables by 99999999 and that of categorical variables by ‘MISSING’. It converts all the categorical variables to uppercase and strips the white spaces present in cases of string.
* It normalizes all the continuous (numerical) variables of the dataset as per the user’s choice.
* It creates 100 estimators by default and finally gives the result based on the mode of the classes.
* It performs hyper parameter tuning to determine the best combination of parameters which would provide a best fit to the model. The user can opt if he/she wants to perform hyper parameter tuning for the specified variables. The hyperparameters to be tuned for this algorithm are “max\_depth” and “n\_estimators”, the pool of values for which are [3,5,7,9] and [50,200,400,600] respectively by default.
* It fits Random Forest model on the training data, sorts the probabilities in descending order and divide the data into 10 deciles. It also scores the Out Of Sample, Out Of Time and Pre-deployment data, sorts the probabilities in descending order and divides the data into 10 deciles using the minimum and maximum probabilities of decile as lower and upper limit.
* It calculates the univariate distribution (count,min,P0.5,P1,P2.5,P5,P10,P20,P25,P30,P40,P50,P60,P70,P75,P80,P90,P97.5,P99,max,missing%) for each dataset.
* It calculates the feature importance of each variable and sorts the variables in their decreasing order of importance.
* It calculates the correlation co-efficient among all independent variables.
* It scores the PSI data, sorts the probabilities in descending order & calculates the PSI value using the distribution of bands on development data and PSI data.
* It saves all of the above mentioned results in an excel file along with the order of variables which would be used to score new data in future (ORDER SHOULD NOT BE CHANGED). These things are saved at the location provided by the user with the project name (also provided by the user), without any human intervention.
* It also saves the AUC chart created for all time periods (in the excel file and as a separate .png file) and the trained model in serialized form for future use in the same location without any human intervention.
* All the files in the provided location are saved such that the file names are followed by the time of the system at which the program is executed. This is done to avoid overwriting of the files when the program is executed multiple times.
* Attached below is a sample output which the code would generate.

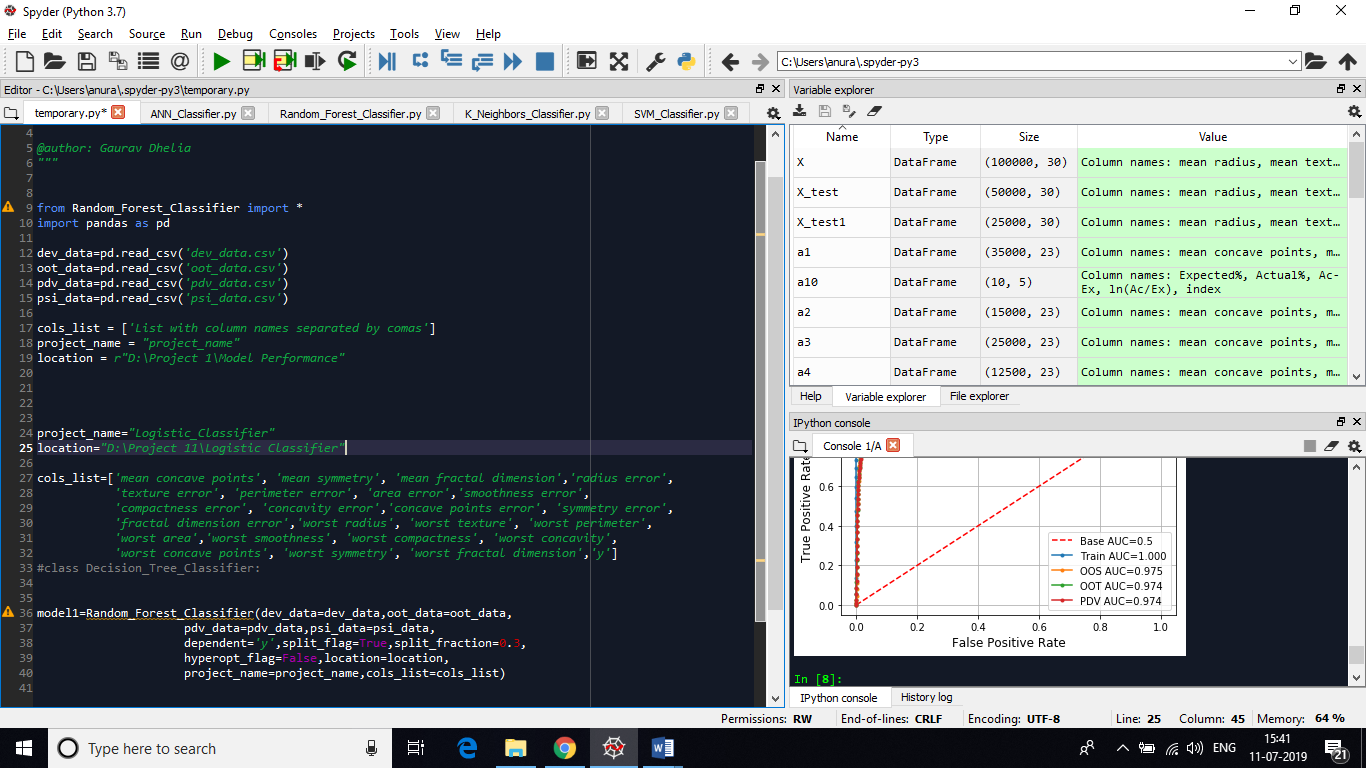
Steps for using the Random Forest Classifier module:-

* Open any Integrated Development Environment (IDE) to run python codes. Eg:- Spyder, Jupyter Notebook, etc.
* Set your working directory. Working directory is where all your data and codes reside.

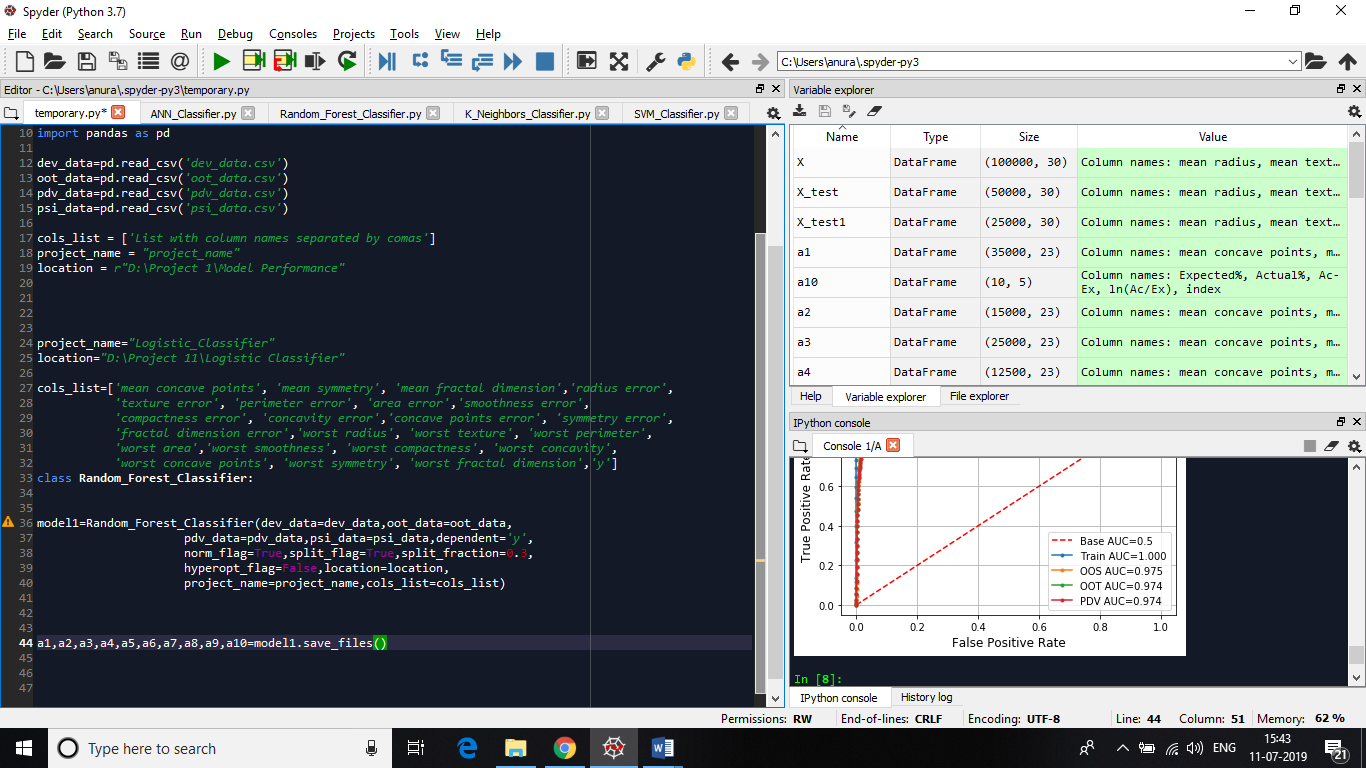


You need to specify your working directory in quotes as shown above.

* Replace F:\User\_Folder\Model Building with your working directory
* Copy Random\_Forest\_Classifier.py in your working directory folder.
* Import the Random\_Forest\_Classifier module. Enter the names of all the .csv files containing the data of different time frames- Out of Sample (if any), Out of Time, Pre deployment and Post Stability, and specify your project name, location where you want your results to be saved and the list of the column names for which the model has to be developed, as shown.



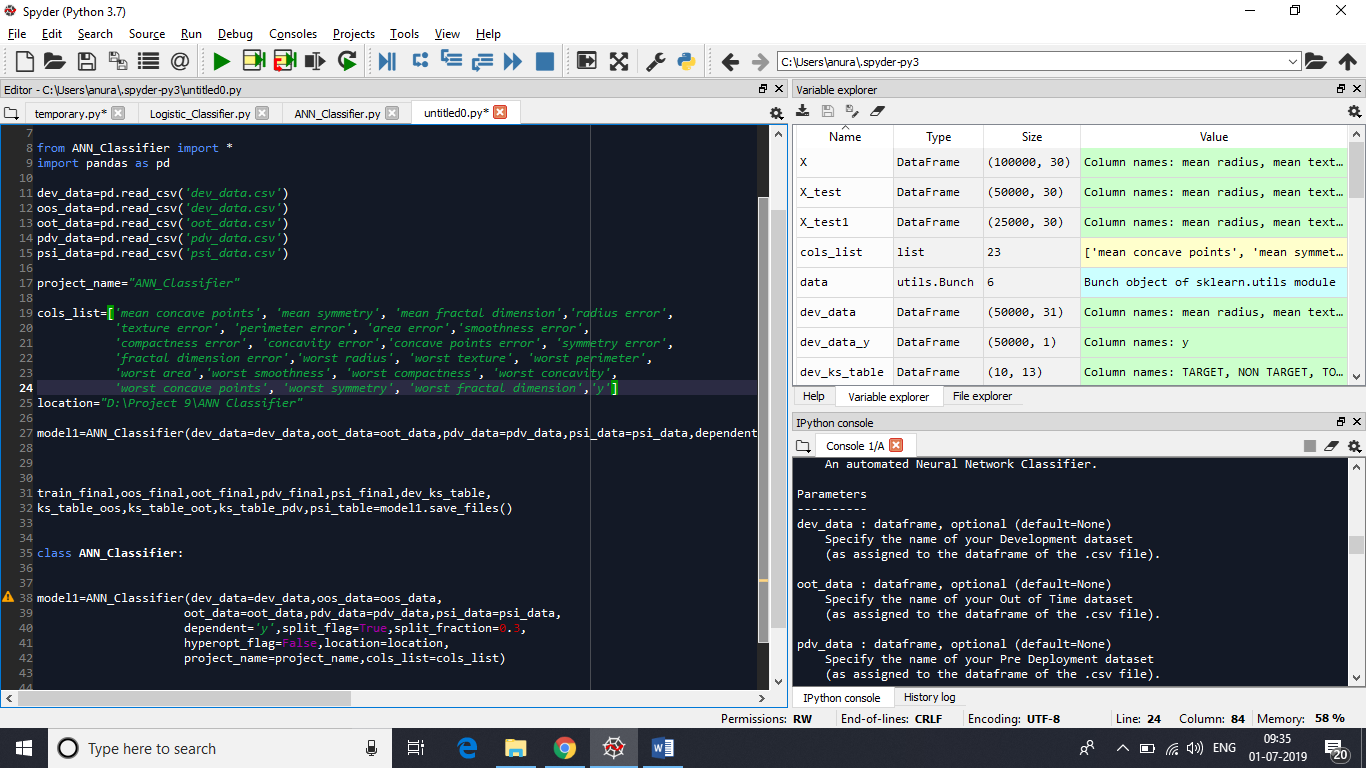
* Create an instance of class Random\_Forest\_Classifier and assign it to a variable (model1) as shown below):



It takes the following inputs:

1. dev\_data:- Specify the name of your Development dataset (as assigned to the dataframe of the .csv file)
2. oos\_data:- Specify the name of your Out Of Sample dataset (as assigned to the dataframe of the .csv file)
3. oot\_data:- Specify the name of your Out Of Time dataset (as assigned to the dataframe of the .csv file)
4. pdv\_data:- Specify the name of your Pre Deployment dataset (as assigned to the dataframe of the .csv file)
5. psi\_data:- Specify the name of your Population Stability Index dataset (as assigned to the dataframe of the .csv file)
6. dependent:- Specify the name of your target variable (dependent) in single/double quotes.
7. norm\_flag:- Boolean value to specify whether you want to normalize the continuous variables of your data. If True, it uses MinMaxScaler() to normalize all the continuous variables and brings them in the range of 0-1. If False, it develops the model on the original values of the data. It is set to False by default.
8. split\_flag:- Boolean value to specify whether you want to split your development data into development and OOS data. If True it splits the data into development and OOS using stratified sampling. If False, it uses the entire development data without splitting. It is set to False by default.
9. split\_fraction:- Float value to specify the fraction of split of the OOS data. It's value is passed only if the split\_flag is set to be True.
10. hyperopt\_flag:- Boolean value to specify whether you want to tune your hyperparameters. If True it uses Grid Search CV to find the best parameters. If False it uses the default parameters. It is set to False by default.
11. location:- The path of the folder where you want your results to be saved, entered in between single/double quotes.
12. project\_name:- Specify the name of your project as mentioned in previous step.
13. cols\_list:-Specify the name of your columns name list created while importing the Random\_Forest\_Classifier module. This column list also contains the name of the dependent variable.

* Using the class instance (model1 as defined above), call the save\_files() function as shown below:



By calling this function, the complete process of developing the model and saving of all the reports in the specified location is performed at once.

* The save\_files() function returns the following results:

1. train\_final:- The final train data after all the processing has been done to it.
2. oos\_final:- The final oos data after all the processing has been done to it.
3. oot\_final:- The final oot data after all the processing has been done to it.
4. pdv\_final:- The final pdv data after all the processing has been done to it.
5. psi\_final:- The final psi data after all the processing has been done to it.
6. dev\_ks\_table:- The KS table created on the training data after sorting all probabilities in descending order.
7. ks\_table\_oos:- The KS table created for the oos data using the ‘dev\_ks\_table’.
8. ks\_table\_oot:- The KS table created for the oot data using the ‘dev\_ks\_table’.
9. ks\_table\_pdv:- The KS table created for the pdv data using the ‘dev\_ks\_table’.
10. psi\_table:- The table created on the psi data showing the calculation for the psi.

* For more information to the class Random\_Forest\_Classifier, the user can look at the ‘Docstring’ after executing the program, which contains a description of all the functions contained within the class. It can be accessed by typing the following as shown, in the IPython Console of Spyder.

