AMOD 5610 Project Prototype

The nearly complete project prototype should have all key features functional, but final analysis will not be done (and may take some time). At this stage most of the project should work, and testing to verify the correctness of implementation will be done.

**Name – Sahil Nagpal (0670042)**

**Name – Pramina Patil (0663967)**  
  
  
Prototype

Explain what of your project works at this stage, what pieces are functional or if there any key functional elements which currently don’t work

|  |
| --- |
| **Finding a Business Problem** (**COMPLETED**)- There is a high degree of competition in today's world, which raises employee pressure. High competition leads to unfulfilled expectations that cause health problems for an employee, and health problems can lead to employee absenteeism. The absence of the resource (purposely employees) may now lead to delinquency of the imperative deadlines which may deficiently affect the business models and the business profits as well. Finding a business problem in today’s era is quite a difficult task to work upon since finding the problem which needs a deliberate solution is become a part and parcel for development.  **Loading the Data (COMPLETED)** – Gathering the right information (raw data) from a different source is quite a managing task which needs attention and some pre requisite knowledge. Scrutinize the different data from different sources was a learning task. Once we capture the right format and the right informative data , we will load the data for further pre-processing and to produce informative yields. We will capture the data either in the comma separated format or in the text format and load it using the python libraries.  **Identify the Independent Variable** (**COMPLETED)**- It is more complex than it first appears to decide which variable is the most significant. In this phase we will finalize which one variable is most important. Although there is no one -size fit all definition of the important variable. From this part we will have our independent variable ‘Y’.  **Determine the Algorithm to develop Model** (**COMPLETED**)– This is one of the most essential segment of producing the outcome. In this phase we will catch the algorithm which fits well to our use case as well as to our corresponding data.  **Data Pre-processing** (**COMPLETED** and **TESTED)** - We typically think of big datasets with a massive number of rows and columns when talking about data. Machines do not understand free text, picture or video information as it is, 1s and 0s are understood by them. So, if we put all our photos on a slideshow and expect our machine learning model to be trained only by that, it probably won't be good enough!   * Remove or Treat Missing Values. * Remove Outliers * Standardize the data. * Normalize and Scale the data   **Split Data into Train and Test** (**COMPLETED** and **TESTED**)- Divide our train and test data and build the model on the set of train data.   * Apply the Algorithm – Determine the respective algorithm and apply that over here. * Train the Model - We will first segment the data into the train and the test. Our model will passed on a train data set. * Test the Model - Then we need to test with a new data set that tests data sets when we have successfully established our model.   **Validate the model** (**COMPLETED** ,**TESTED** and **FUNCTIONAL**)– In this phase we will figure out the generalization ability of the trained model.   * Find the intercepts and coefficients. * Interpreting the coefficients.   **Save the Model** (**COMPLETED** and **TESTED**)– In this point of solving phase , we will save the model and deploy it on the cloud for its effective usage. In order to save the model we need to pickle the model. We will also create the executable file for the same.  **Model Performance and Sanity Testing** (**COMPLETED** and **TESTED**)– We will test the model on the new data set with same feature and will also perform the sanity testing of the result generated from the new dataset.  **Result Presentation** (**COMPLETED** ,**TESTED** and **FUNCTIONAL**)– It is known to be a best practice to store the data and the predictive results under a certain database. In our case we will either take the MySQL or the PostgreSQL to store this respective information. We will then make a connection between the database and the Tableau (since it is an interactive and strong tool to present the data in more happening way) and present the results. |

# Installation

Explain how to install/setup/run your project

|  |
| --- |
| In order to use the model or to run the project **we have created the model and the scale file**. Please follow the following steps for fruitful and the proper formatted result. For the sake of the usage, we have created the scale and model file using the pythons pickle module (we have created the **model deployment files** which is being used for creating the API’s , Django and Flask Web Sites).   1. Open any python editor (Integrated Development Environment) or you can also use the Jupyter Notebook. To the customer centric deployment , the results can be generated in just four lines of code. 2. With this assignment, We will share a zip file which consist of:  * absenteeism\_module.py (This python file consist of the methods and functions which will process and executes the input data). * Absenteeism\_new\_data.csv (This is the input file which a user or a client will bring to use the model). * Model (This model is the COCOM type executable model which generates the probability of the Absenteeism). * Scaler (The scaler file act as the input file for any type of method and this is generated using the pickle module).  1. Once you open the Jupyter notebook , Import all the module from absenteeism\_module.py      1. Now , create a variable of your choice and import the model and scale under the absenteeism model method.      1. And now , Import the input data (the input data can be raw data with outliers and all other error prone components).      1. Now the fourth and the final step (yes it’s the last step), use your variable which you have used for storing the model and scale values. We have a dedicated method written in for generating the probability output and store your results in a csv for visualization and other purpose.     Note : You can also copy and paste (use) the same code for usage as well. These four lines of code can generate the fruitful results.  Heads Up : You schema must match with the model schema. |

Testing of Functionality  
  
Explain how you are verifying that your project is working correctly at this stage? Essentially, how do you know it’s doing what you want it to do?

|  |
| --- |
| For the verification purpose we have used few statistical technique and we have also used some of the unit test cases which helps us to verify the raw data and the pre-processed data.  **K-Fold Cross Validation**  A statistical approach used to estimate the ability of machine learning models is cross-validation. Since we are using the specific algorithm to build the model , we will also build the pseudo model using the other algorithm and compare the results of the pseudo models with the actual implemented model. For our model, we have checked the random probabilities of each row by cross validating the probability values.  **Classification Metrics**  The classification metrics also known as the confusion matrix. One of the most intuitive and simplest (unless, of course, you are not confused) metrics used to find the model's correctness and accuracy is the Confusion matrix. We also consider the **Accuracy Score** , **Precision score** and the **recall values** in order to verify if the model is working as per the expectations.      **Absolute Testing of the functionality – Checking the Expectations**  For the absenteeism prediction, we were actually **calculating the probability** of the absence based on the certain parameters. After creating and training the model we have those parameters from the model and it states the absence based on the certain conditions and reason.    Moreover, we have also cross checked the probability of the similar types of parameters by clustering them into same group. We measure the comparable probability which gives a concrete assembled result which can proves that the model is yielding the fruitful results. We have also write unit test cases which asserts if the resultant outputs are actual probability or just the pseudo values. |