AMOD 5610H: Project Plan

Name – Sahil Nagpal (0670042)

Partner Name – Pramina Patil (0663967)

Building on their technical feasibility, students will develop a project plan and schedule, where they layout milestones and plan out all parts of the project, including who will be responsible for which parts, this will be 5-7 pages including diagrams and approximately 1500 words.

# Key project stages

These are not project management phases, rather it is specific to your project.   
E.g.

1. loading data
2. Implement machine learning
3. Training model
4. Testing on real data
5. Validating model
6. Analysis
7. Presentation

|  |
| --- |
| **Finding a Business Problem** - 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** – 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** - 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** – 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** - 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** - 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** – 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** – 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** – 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** – 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. |

Validation and Verification  
  
Verification: How will you know the project works correctly

|  |
| --- |
| Verification is a procedure or methodology used by developers for quality assurance, whereby an inspection of a part, product or service is performed at the end of a step or project to verify or confirm that it meets all the criteria of the regulations or specifications. However when it comes to data science , nobody concretely explains how the entire process will go on.  But for the verification we will follow the **OSEMN Framework** which stands for OBTAIN , SCRUB,EXPLORE,MODEL,INTERPRET.    [1]  This framework helps us to collate and to orchestrate the steps we have designed to execute. After execution of each step in the OSEMN framework , we will verify the results with the actual results to the expected results. For the same , we can also takes the help of the python unit test cases for some couple of executions for instance – we can check if the data is loaded successfully or not.    **Test-Train Split**  Also for the validation , we will also perform the train and the test slit of the data and will examine how the model which is generated for the project will behave when passed through the unseen data. However this method can cause the randomization error which is known as the Sampling Bias (symmetric error).  **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. Comparing and selecting a model for a given predictive modelling problem is widely used in applied machine learning because it is simple to understand, easy to implement, and results in ability estimates that typically have a lower bias than other approaches.  **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 can also consider the **Accuracy Score** , **Precision score** and the **recall values** in order to verify if the project is working as per the expectations. |

Validation: How will you know the project solves the problem you’re trying to solve?

|  |
| --- |
| The validation process normally takes place as one step or several phases of the process , usually in stages, and is used to assess if the project meets the requirements. Validation is the tool for determining the phases of the project.    To make a concrete verification of the steps and the outcome we want , we will verify each and every outcome of each step during the project building phase. Each phase will pass through some sort of unit test cases or some evaluation which will tangible the accuracy of the yield we will be having from project.  To set the bars bit high and to check if the problem is being is solved after the certain outcomes and testing , we will more focus on the deliverables after the model training and check the result of the model with the actual result that should come. A slight change in the processing make a big impact in the outcome which will help us to fix the change.  We will also set some criteria which will eventually help us to find if the project is actually solving the problem or not and those criteria’s are as follows :   1. We must verify if ***Deliverables == Expected*** ***Result*** after each and every step. 2. We will pass the result from the mathematical validation (Cross Validation , Accuracy Score , Recall Values) and if the values are doubtful or not producing satisfactory figures , then the model will be again revaluated for numerous until the results are above expectations. |

Schedule, including individual responsibilities  
  
Based on the Key phases in part one, lay out who is doing what when.   
  
If you know how to do a Gantt chart you can do that (though keep in mind that in the real word, project management tools like Gantt charts, Scrum boards, and Kanban boards are usually fiction done for managers who are incapable of knowing better, and technical people build first, and write the plan for how they supposedly did it after the fact).

|  |
| --- |
| Schedule and dividing the task will definitely make any project work and execute more faster. In our project , we are following the principle of the **Agile Methodologies**. The three major principles we are following in order to make the progress in our in our project are as follows :   1. Individual and Collaborations over the execution models and the tools. 2. Working model over comprehensive requirements and documentation. 3. Responding to the new changes in the execution.   The **Agile framework** we are using over in this project is the **Kanban**  Kanban acts as a fan-in and fan-out mechanism where the team pulls user storeys from an intake board and funnels them before they are done through a staged creation process.    The Kanban board we have created - <https://researchinmotion.kanbantool.com/b/653672-absenteeism-project#?>  In this Kanban board we have created the project name and under that project we have distinguish and allotted our respective task with the **subtask** , **deadline** (time) and the **severity** of the task.    Every project has a backlog of tasks to get done, and a series of process states that a task must follow and go through before it is delivered. Using the Kanban board, both of the team members can instantly see how tasks are moving through the process and what needs to be worked on.  Kanban also lets us to track our cycle time and throughput consistently, displaying us how our productivity changes over time. The faster the tasks can move through your process, the more tasks we can get completed as a team.  As of now we have added the task up to the preprocessing tasks. We have also created a separate section for the task which has not been completed under the certain deadline and moved to the **Backlog Section**. As we complete our task we will moved to the **Done Section** and once we start making any progress in any task then we will drag it to the **In Progress Section**.    As shown above, every time we receive a project task we follow the process to work on it:   * Identifying different project tasks and prioritizing them as per the importance. * Conducting a scrum meeting and discussing various aspects of the tasks like time to be allotted for completion of the task, what is the scope of the task and story point and how to overcome any foreseeable complication. * Moving forward we add the task to the **To-do** list and work on it as per the priority we have agreed on. After all the high priority tasks have been worked upon and moved to the **Done** list, we begin work on the next tasks and add it to the **In Progress** Tab. * Finally, once the task is completed, it is marked as **Done**.   This helps us to keep a track of all the tasks that have to be performed and how much efforts are required for each of the task. This makes time management as a time efficient. |

**Reference**

[1] Lau, D. (2019, January 10). 5 Steps of a Data Science Project Lifecycle. Retrieved October 21, 2020, from <https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492>