Capstone ML Project

# Session: 4- December, 2021

## Feedback:

1. Project Plan (Grantt chart) looks good but it is very detailed. Need to find a way to compress it such that it will fit in the final report document.
   1. Ranga is working on this. (Done)
2. Flowchart should start with a problem statement, and we need to reduce the decision boxes such that it is more streamlined.
   1. Indra needs to look into this.
3. Running late on Report (document), should have started with Chapters (Table of contents) to get an idea of the report.
   1. Different sections of the document, (to be discussed after the meeting)
   2. We are starting with the final report format itself contents keep getting added on.
   3. One owner for the document.
4. Should have merge the code into master copy and then presented to Mentor. (This is was missed)
   1. Do a live meeting and plug in the code into the master copy.
5. Data interpretation went well (Shoutout to Ranga), but it is essential that all team members, know about various parameters in the data set and their co-relation when it comes to prediction of price of the house.
   1. Data representation of the zip codes to identify which particular area is expensive.
6. Problem statement needs to be formulated based on the given dataset and not be copied from the PDF provided.
   1. Creation of Intelligent Regression based data model to predict house / home prices on basis of sales data in Seattle region from 2014 to 2015. Data models take into account various features like area, location, amenities and condition. (Finalized and agreed)
      1. We can add a point in the document on our rationale to drop various columns, drop dayhours, cid and other parameters.
      2. Improvements that can be done, if required for different regions
         1. first run classifiers to separate out regions
         2. Run regression on different datasets for particular region.
7. EDA with visualizations was accepted (Sunil’s presentation) but we need to have selected graph or charts snippets in the master copy as this would increase the length of Jupyter notebook.
   1. Keep charts based on target variable prediction only and not include all univariate and bivariate analysis.
   2. Heat map should be present.
8. Team should have discussed by now on different algorithm comparison, this should be done before next session.
9. Explore Outlier detection Library PyOD and try to use in our project (if possible).
10. Coding Guidelines:
    1. All libraries should be imported at the start of the file or Jupyter notebook.
    2. Proper comments on each code block which would explain what exactly are we trying to do here.
    3. Explanatory statements should be present which would describe about the inference from the output of above code block
    4. Columns such as “dayhours” need to be converted to timestamps such that data type changes from object to Integer. (If we are dropping dayhours, not required)

## Deliverables for Next week

1. Structure Project Report with Content including: (Interim Report)
   1. Snippets from Jupyter notebook
   2. Inferences derived by the team
   3. Project plan,
   4. Flowchart
   5. Complete Table of contents describing the whole structure of the document.
2. Merged code with complete EDA, Baseline linear regression and model comparisons.
3. Explore on Model Score Boosting techniques that have been taught.