Meeting Record:

Meeting 1

[Feb 15, 2021 07:00 PM]

- Discussion on the assignment requirement.
- Agreed actions: Dataset suggestion list.

Meeting 2

[Feb 18, 2021 07:00 PM]

- Review suggested dataset list.
- Created a list containing 6 datasets.
- Agreed actions: Review the 6 listed dataset and vote for one.

Meeting 3

[Feb 22, 2021 07:00 PM]

- Decided to go for YouTube Dataset

Meeting 4

[Feb 28, 2021 07:00 PM]

- Discussion on the work plan
- Agreed actions: Start with Research

Meeting 5

[March 9, 2021 12:30 PM]

- Not sure about the Problem Statement
- Need to go back and look for new dataset
  - Jane Street Market Prediction
    https://www.kaggle.com/c/jane-street-market-prediction/overview/evaluation
  - Human Protein Atlas - Single Cell Image Classification
    https://www.kaggle.com/c/hpa-single-cell-image-classification/data
  - Youtube API
- Agreed actions: New Problem statement and Check for new Dataset

Meeting 6

[March 13, 2021 07:00 PM]

Agreed Action before tomorrow meeting

1. Please bring dataset suggestion and problem statement
2. Go through dataset and notebook shared by Aniruddh. So that we can point out any issues right in tomorrow meeting.

Meeting Agenda

1. Discuss on the blackfriday dataset proposed by Aniruddh
2. If team members have found any other dataset and problem statement. Lets discuss on that.
3. Make a list of task and divide
4. Update the ppt
5. Maybe create a GitHub for the assignment to add the codes and documentation.

Agreed actions

1. Literature Review on shopping sell **- Aniruddh, Gelmis and Tenzin**
2. Different Strategy to impute missing data in product_category_2 **- Shubham**
3. Feature Engineering **- Shubam**
4. Model Implementation **- Aniruddh, Gelmis and Tenzin**
5. Documentation **- Tenzin**

Meeting 7

[March 20, 2021 07:00 PM]

Missing value mechanism (MCR), imputation method, Analyze the comparison and probability stuff done on aniruddh Jupiter notebook - **Shubam**
List out Models - **Gelmis**
Read about evaluation metrics - **Aniruddh**
Evaluate the Model based on the metrics above
Documentation - **Tenzin**

Meeting 8

[March 29, 2021 12:00 PM]

Read Literature Review in the document **- Aniruddh, Gelmis, Shubam**
Model Implementation **- Aniruddh, Gelmis, Shubam**
Introduction **- Shubam**
Find few more Literature **- Tenzin**
EDA Colab Notebook **- Tenzin**
Evaluation metrics **- Aniruddh**

Meeting 9

[April 1, 2021 7:00 PM]

Discussions:

Model Implementation Performed by **Aniruddh**: LR, DT, RF
Literature Review, Outlier Detection and EDA Colab notebook updates by **Tenzin**
Introduction progress by **Shubam**

Agreed Actions:

Model Implementation - **Aniruddh**
Literature Review (Sales) and Continue with Documentation - **Tenzin**
Model Implementation - **Gelmis**
Introduction **- Shubam**

Meeting 10

[April 6, 2021 8:00 PM]

https://github.com/aniruddh1804/Data-Science-Projects/blob/main/Black%20friday.ipynb

Model	Evaluation			
	MAE	R-squared	F-Test	RMSE
Linear Regression				
Polynomial Feature Transformation (LR)				
Decision Tree Regression				
Random Forest Regression				
XGBoost				

- Calculate multicollinearity, may be using chi-square test
- Implement ridge, lasso regression to eliminate non-essential variables
- Implement XGBoost for regression analysis
- May be convert age to ordinal variable, and see the impact on each of the models
- Prepare a table for each variation of features and models, for both RMSE and adjusted r-squared

Evaluation measures used for regression analysis:

1. Mean Absolute Error (MAE) / Mean squared error - used to indicate how much is the overall error in our dataset
2. r-squared - used to explain the variation in the dependent variable explained by the independent variable
3. adjusted r-squared - takes into account the number of predictors
4. F-test - used to see how significant are the predictors in comparison to an intercept only model (ideally should be much less than 5 %)
5. p-values of each of the coefficients in the regression - to test their statistical significance
6. Omnibus, skew and kurtosis of the errors - to test if they are normally distributed and what are the chances of errors being normally distributed

Documentation:
Null Value handled - Prob Distribution (Justification)

1. Finish literature review (Tenzin)

2. FIgure out business understanding, data preparation and data understanding and update google doc as well (Shubham) - this would contain missing values, multi-collinearity, distributions, unique values, handling missing values
3. Read literature review, and implement models based on metrics I've given, and will mention (Gelmis)

Meeting 11

1. Read Shubham's introduction, suggest changes - Aniruddh, Tenzin, Gelmis
2. Proof read related work (think of points to criticize) and dataset part
3. Business understanding, data understanding, data preparation (Tenzin, Shubham)
4. Proposed Methodology (Tenzin, Shubham)
5. Conclusions - Gelmis, Aniruddh

Meeting 12

1. Analyse what Aniruddh has done - chi-squared, AIC, BIC, and dropping product category 2 (based on p-values of each column after OLS regression) and seeing results -Tenzin, Shubham
2. Implement lasso regression, ridge regression, or elastic net regression - Aniruddh
3. Proofread it, tidy up the report, model and evaluation - Gelmis
4. Video - Tidy up the notebook, and the presentation - Gelmis