Paper Outline Planning

Thesis Type: **Empirical**

Introduction and history on roles of data imputation

The first section of the paper should play a role to justify the relevance of the material, which in this case is data imputation on missing values. The introduction would explain the history of data analytics in general into the direction of handling missing values. The aim is to follow the trails of footprints that the data analytics field has had in the past, and explain not only the importance of data imputation itself (why it is crucial to have), but also the methods of execution. (why it is crucial to do it right.)

Exposure to Various Imputation Methods

This part of the paper introduces the various methods of imputation taht are available to pick from, and its algorithmic/mathematical structure. Some of the members that will have its name in this section are quite trivial (ex. deleting datapoint, mean, forward fill, etc.), and does require less explanation. However, there are some methods that would be more complex to explain and go through, especially ones that involve complex algorithms like deep learning.

Empirical Section

In this section, multiple datasets (at least 2, for regression and classification) to demonstrate the effects and consequences of the various imputation methods. The "experiment" will be conducted in the following manner:

1. Take a dataset from a famous Data analytics/ML competition sites such as Kaggle. (Assume there are no missing values in the original dataset)
2. Recognize the algorithm and its set of hyperparameters that the most accurate competitor has utilized.
3. Create multiple artificially damaged datasets based on some policy (The random method is something that could make sense) [1 -> n] datasets
4. Impute the missing values on various imputation methods mentioned in previous sections, and run the ML algorithms on the imputed dataset. Record the accuracies.
5. Conduct 3,4 multiple times (30+ for the sake of ease of experimentation due to the central limit theorem) to conduct a statistical hypothesis testing, to eventually find a statistically significant (inferior, or that the original accuracy being "statistically significant to be superior") imputation method against the accuracy with dataset w/o missing values.
6. Repeat 3-5 multiple times for different "missing rates"

Conduct 1-6 for multiple datasets for both regression and classification.

Result Description

This section is solely for expressing the massive experimentation result obtained in previous section, in various methods. (Table, plotting, etc.)

Hypothesis Testing

Based on sections above, this section is dedicated for recognizing if the imputation methods have created any statistically significant differences. Having the distribution of accuracies obtained under an imputation method, we see if the original accuracy under the full dataset is significantly superior or not.

Result Evaluation under the experiments

Based on the hypothesis testing mentioned above, the paper suggests which method performed the best, worst, everything in between, under the experiment conducted. This is not an evaluation of absolute and global "ranking" of the imputation methods, but rather a mere observation on how some methods yielded different result in reaction to the experiment environment given.

Business Implications and Potential Danger (Potentially fusioned with conclusion)

This section turn more theoretical, on how continuation on researching of imputation methods are crucial, by demonstrating potential changes and inaccuracies in conclusion drawn and some risks that would come with that decision, as a consequence.