P9120 Final Project Proposal

He Jin

hj2479

**Project title:**

Using Random Forest and Gradient Boosting to predict German used cars prices in R

**The motivation for this project:**

Use different statistical methods to predict the German used cars price in German market and compare the test accuracy as well as running time between each method and each package, and eventually select the optimal one to make prediction given the testing data.

**Data Source:**

The dataset I am going to use is from Kaggle used car price database: <https://www.kaggle.com/orgesleka/used-cars-database>

**Planned Analysis Structure:**

1. Data preprocessing and select the target feature we would like to include in our model
2. Basic EDA for the data
3. Fit the data using Random Forest model in different packages
4. Fit the data using Gradient Boosting Method in different packages
5. Make comparison between two different method
6. Select the best method and best package and make conclusion

**Planned Timeline:**

|  |  |
| --- | --- |
| **Time** | **Goal** |
| November 15th – 20th | Wait for the feedback from proposal and conduct data cleaning |
| November 20th – 23th | Select the target feature and conduct basic EDA for the dataset |
| November 23th – 30th | Fit the Model with two different methods |
| December 1th – 6th | Summarize the results and prepare for the presentation |
| December 7th – 18th | Finish the report |

**Additional notes:**

Actually, this is the second time I would use this German car used database. Last time, I was using it as the dataset for the final project of data science 2 project. However, I only made a simple comparison between some of the statistical learning methods. Here, I would like to make a detailed comparison specific between Random Forest Regression and Gradient Boosting method since they are always with each other in the machine learning field and I am wondering how they would perform in this specific used car dataset.