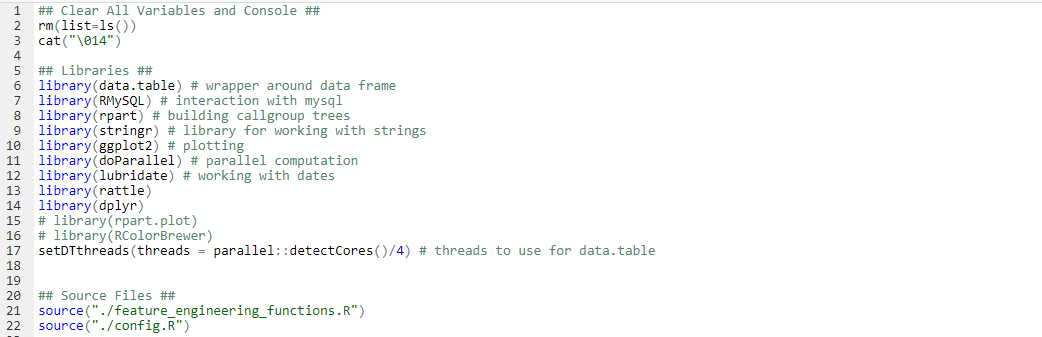
**Feature Engineering Task 2 Guide**

**Data Preparation**

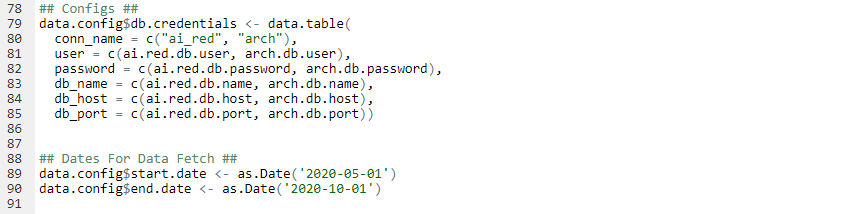
1. **Pre Reqs**Clear all the environment variables and console  
   Load the libraries  
   Load the source files



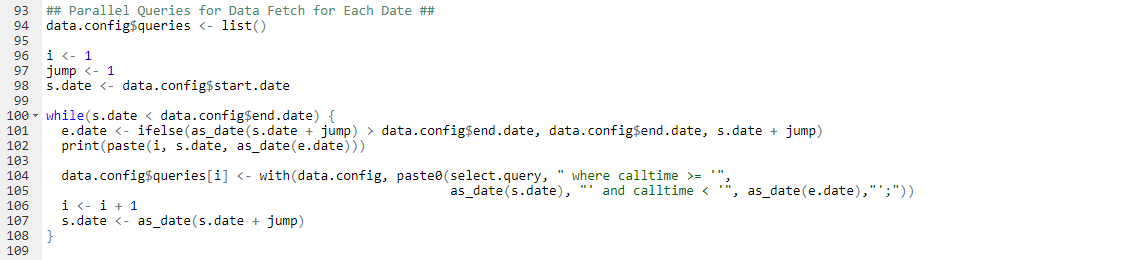
1. **Input the Query for Data fetch from SME**Add in the target metric along with the testing variables from the dataset.



1. **Loading the DB credentials and setting the dates for data fetch**



1. **Add in the filtering criteria if required apart from the calltime**

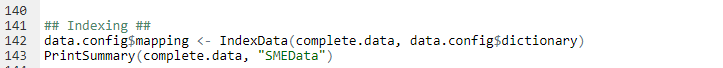


1. **Create a data dictionary for the data and fetch the data**set the relevant columns that you want to test as is\_callgroup.  
   \*\*You can also change the data types of the numerical fields to numeric so that it’s not indexed.  
   \*\*the smalls threshold is set to 0.05 in this case and is only being applied to the columns which are indexed as we don’t need to index the columns that are already numeric.



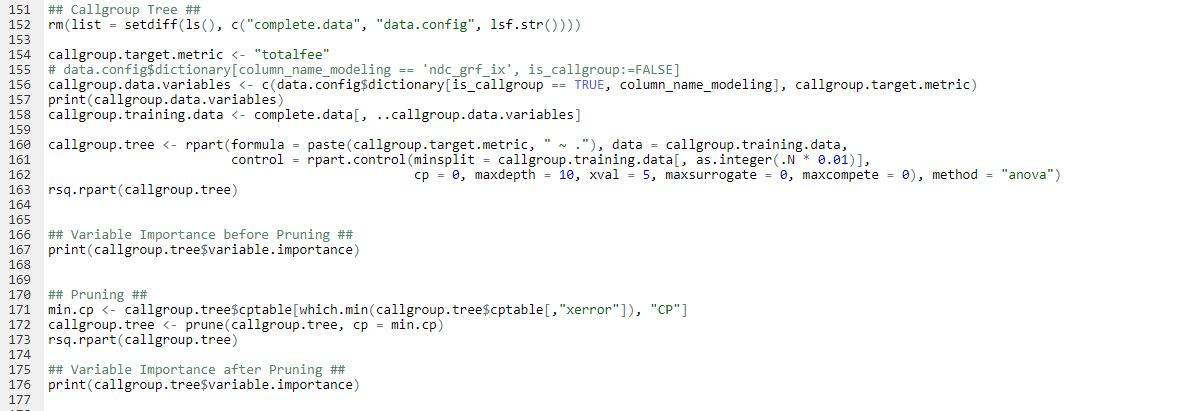
**Index Mapping**

1. **Create a mapping for indexing which will be used later in the script**



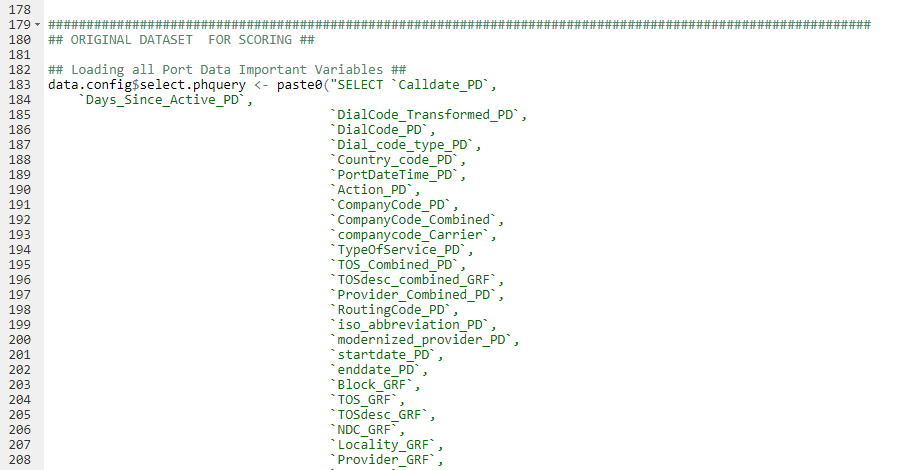
**Tree Building**

1. **Create the Callgroup Tree**Set the target metric before building the tree and prune to get the final output.

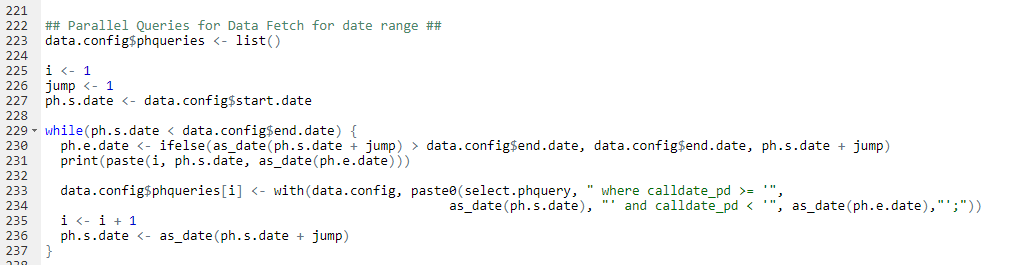


**Preparing the Original Dataset for Scoring**

1. **Input the Query to fetch the original dataset**



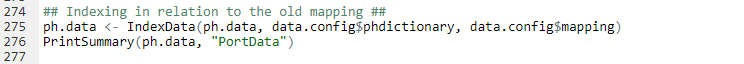
1. **Create parallel queries using dates in the original data**\*\*If there is no date column in the original dataset, you can also create separate queries using btn or any numerical variable.



1. **Create a Data Dictionary for the original Dataset and fetch the data**

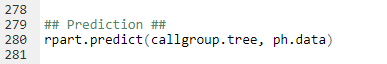


1. **Apply the index mapping from the previous dataset onto the original Dataset**

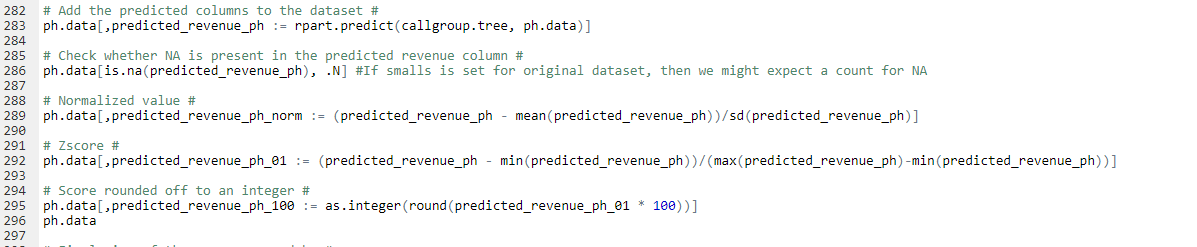


**Predicting the Original Dataset**

1. **Use rpart.predict to predict the original dataset using the callgroup tree**\*\*If rpart.predict does not work, you can also use predict

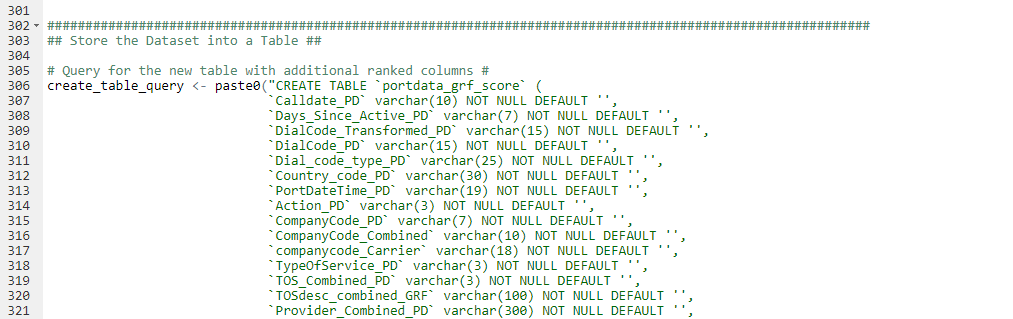


1. **Add new columns for predicted score to the original dataset**

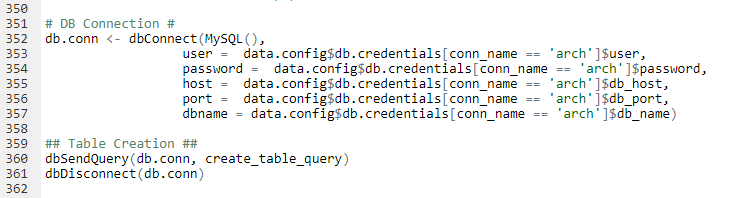


**Loading the dataset back to a MySql Table**

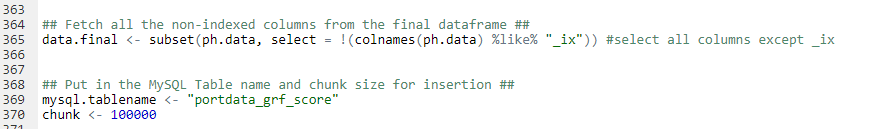
1. **Input the Query for table creation**\*\*this create statement will be the same as of the original table, we will just need to add the scoring columns along with it.



1. **Create the table**



1. **Fetch all the non-indexed columns**\*\*Columns with \_ix are not required so it is preferable to exclude them from the data that will be loaded in the final table.  
   \*\*Input the table name (mysql.tablename) where data is to be inserted   
   \*\*Set the chunk size to divide the data into chunks for parallel insertion(this is arbitrary and will be dependent on the size of the dataset)



1. **Chunking and Function for parallel insertion**\*\*No need to change anything here.



1. **Insertion of data to MySQL Table**

