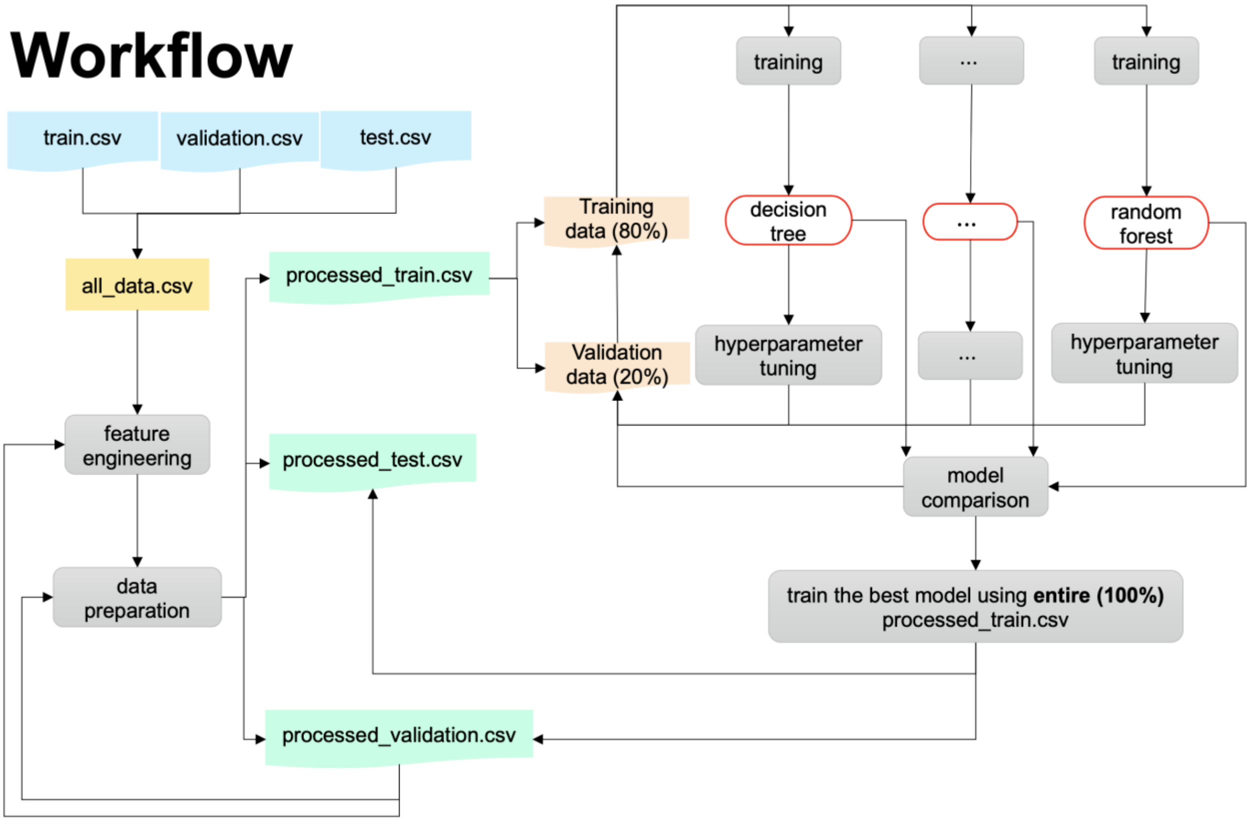
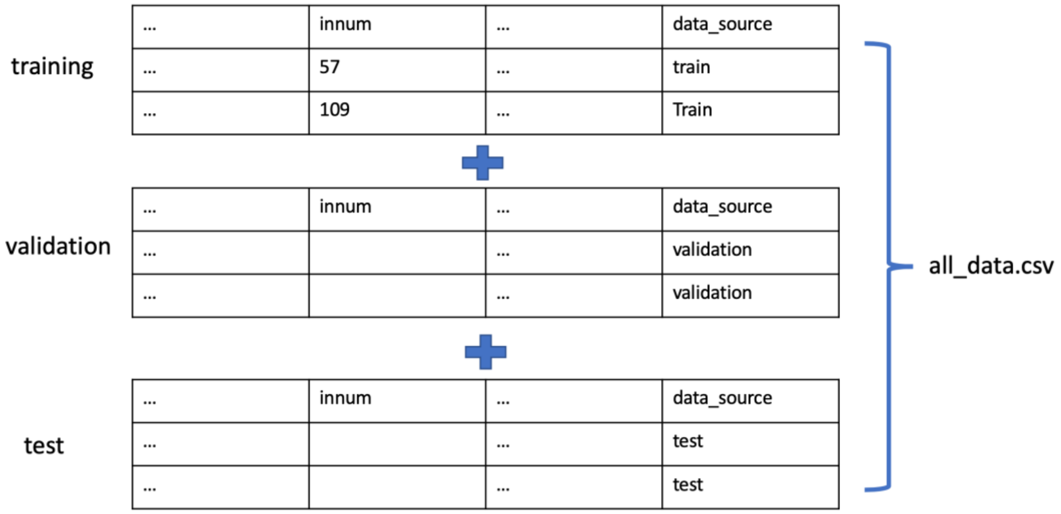
Case Competition Workflow



all_data.csv combined observations in train.csv, validation.csv, and test.csv. We start with combined data to do feature engineering and preprocessing so that all observations follow the same strategies.



training

...	innum	...	data_source
...	57	...	train
...	109	...	Train

validation

...	innum	...	data_source
...		...	validation
...		...	validation

test

...	innum	...	data_source
...		...	test
...		...	test

all_data.csv

After that, you can split the processed combined data back into processed train.csv, processed validation.csv, and processed test.csv using the column 'data_source'. Use processed train.csv for model training and parameter tuning. If you want to get feedback on validation set, then make predictions on processed validation.csv. For final submission, make predictions on processed test.csv.

Detailed instructions:

Step 1. Download ***all_data.csv***

Step 2. Feature engineering – Excel

- This step is optional
- Generate new features
- Remove features

Step 3. Preprocessing - Rattle

1. Load the file into rattle
2. Do not partition data (uncheck partition)
3. Select features you want to use
    a. Mark the features that you don't want to use as ignored
    b. Do not ignore id, innum, and data source
4. Recognize correct data type for the selected input features
    a. Categorical features with incorrect data type
        i. Transform -> recode -> As Categoric
5. Missing data imputation on selected input features
    a. Numeric
    b. Categorical
    c. Need to check if missing values in one feature truly means those values are missing
    d. Do not impute target var: innum
    e. Do not delete observations with target variable value missing
6. Transformation
    a. Numeric
    b. Categorical
        i. Recode as indicator variables
    c. Do not transform id, innum, and data source
7. Deal with ignore features
    a. If your preprocessing is done:
        i. Delete ignored features
8. Export the processed data into a csv file ***'processed_all_data_version1.csv'***

Step 4. Break the processed data into three separate csv files - Excel

1. Apply filter on data_source column to break the ***'processed_all_data_version1.csv'*** into
    a. processed_train_version1.csv
        i. filter all observations with Data source = 'train'
        ii. select them, copy and paste them to a new excel
        iii. save as csv file
    b. process_validation_version1.csv

        i. filter all observations with Data_source= ‘validation’
        ii. select them, copy and paste them to a new excel
        iii. save as a csv file
    c. ***processed_test_version1.csv***
        i. filter all observations with Data_source= ‘test
        ii. select them, copy and paste them to a new excel
        iii. save as a csv file
2. Open ***processed_train_version1.csv***
    a. Delete data_source
3. Open ***processed_validaton_version1.csv*** and ***processed_test_version1.csv***
    a. Delete data_source
    b. Delete innum

Step 5. Modeling - Rattle

1. Load ***processed_train_version1.csv*** into rattle
    a. Partition the data into training data and validation data (80/20/0) using a specific random seed
    b. Check if input features are correctly recognized
    c. Check if ID is marked as ident
    d. Check if target variable is correctly recognized
    e. Make sure target data type is numeric

2. Parameter tuning
    a. Train a model with different parameter settings
    b. Compare performance of different parameter settings on validation data
        i. Evaluate -> Score-> check the specific model->Validation->Identifier
        ii. Open the output file in excel and then compute MAE
        iii. Choose the optimal parameter setting which gives the lowest MAE

3. Comparing models
    a. Train different models with their optimal parameter values
    b. Compare performance of different models on validation data
        i. Evaluate -> Score-> check models you want to compare->Validation->Identifier
        ii. Open the output file in Excel and then compute MAE
        iii. Choose the best model which gives the lowest MAE

4. Retrain the best model use the entire ***processed_train_version1.csv***
    a. Re-partition the data using sizes (100/0/0)
        i. All observations are used for training the best model
    b. Train the best model on the new training data
    c. Save the whole project as ***.Rattle*** file
        i. Give it an informative name (e.g., ***DecisionTree_version1.Rattle***, if your best model is decision tree trained using ***processed_train_version1.csv***)
        ii. Later, you can reuse this model by loading the ***.Rattle*** file

Step 6. Make predictions on the separate ***processed_validation_version1.csv***

1. Evaluate -> Score -> chose the model -> CSV File -> choose the ***processed_validation_version1.csv*** -> Identifier
2. Save the file as Team_XX.csv, where XX is your team number (e.g., Team_12.csv)
3. The output file contains two columns: ID and your predictions (e.g., glm use linear regression)
4. Submit the file on ICON and get feedback from TA.

Step 7. Improve performance

1. Try other models on using current ***processed_train_version1.csv*** (**Step 5**)
2. Apply different feature engineering or preprocessing strategies (**Step 2** or **Step 3**)
   a. Need to start from original ***all_data.csv*** file
   b. Include more features, remove less relevant features or change preprocessing strategies
   c. Output the processed file again (e.g., ***'processed_all_data_version2.csv***)
   d. Break ***processed_all_data_version2.csv*** into 3 separate files again (**Step 4**)
      i. ***processed_train_version2.csv***
      ii. ***processed_validation_version2.csv***
      iii. ***processed_test_version2.csv***
   e. load ***processed_train_version2.csv*** into rattle and repeat Modeling steps to select best model (**Step 5**)
   f. Make predictions on the ***processed_validation_version2.csv*** (**Step 6**)
   g. Get feedback from TA and decide if you want to make further improvement
      i. If you want to make more improvement
         - **Step 7**
      ii. If not
         - Decide your **FINAL** model
         - Load the ***.Rattle*** file associated with your **FINAL** model (e.g. ***DecisionTree_version1.Rattle***)
         - Make predictions on the corresponding ***processed_test_versionZ.csv*** , where Z is the version number corresponding to your **FINAL** model (e.g., ***processed_test_version1.csv***)

Step 8. Make predictions on the separate **processed_test_versionZ.csv** using your **FINAL** model

1. Evaluate -> Score -> chose the model -> CSV File -> choose the ***processed_test_versionZ.csv*** (Z is the version number corresponding to your **FINAL** model) -> Identifier
2. Save the file as Team_XX.csv, where XX is your team number (e.g., Team_12.csv)
3. The output file contains two columns: ID and your predictions (e.g., glm use linear regression)
4. Submit the file on ICON by the deadline