IML2024 Term project Report

Patrik, Mikko Kallio, Tuuli Toivanen-Gripentrog

2024-11-22

Data Exploration

Feature correlations

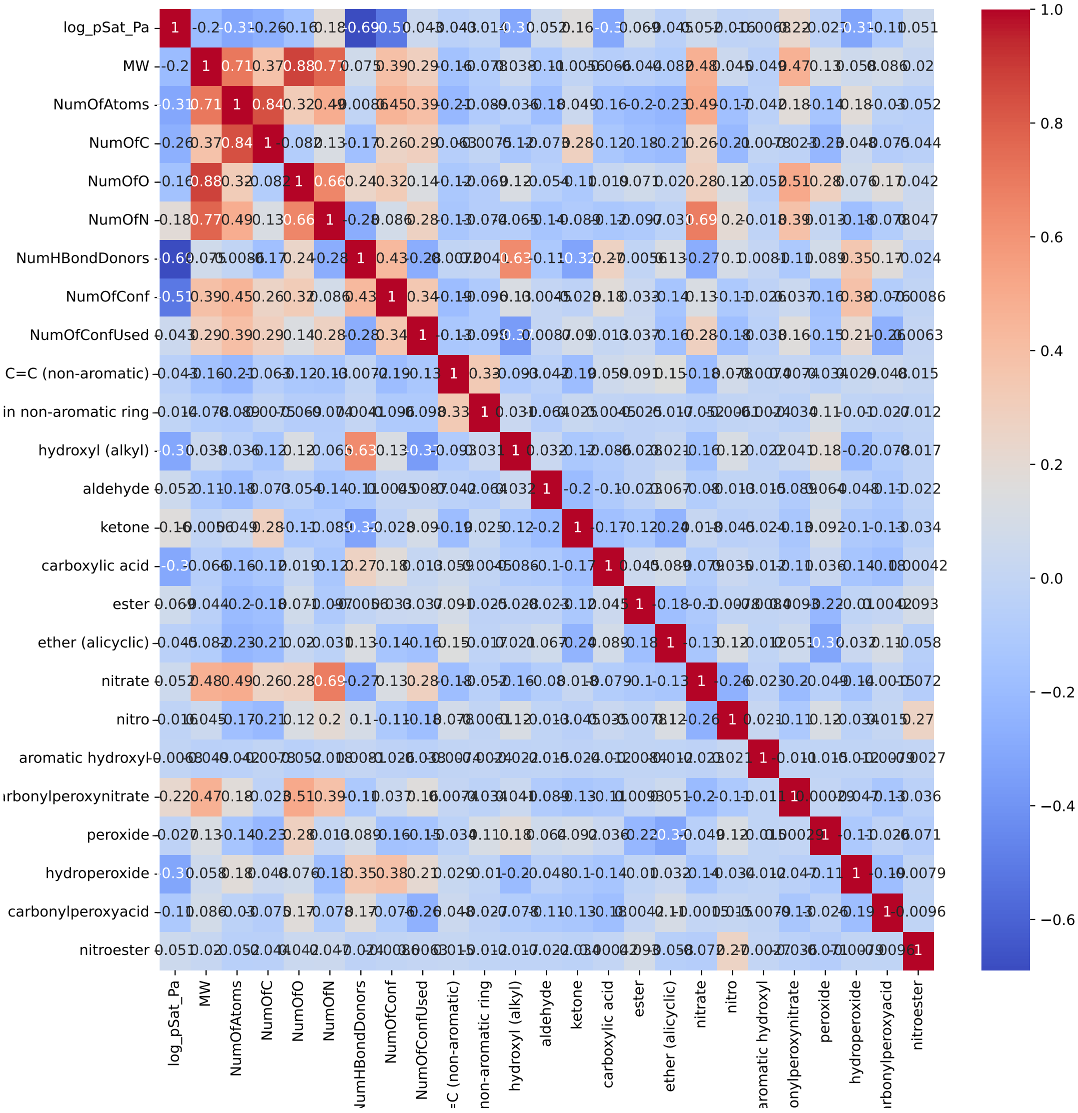
First we analyzed the different features and how they correlate between the target column log_pSat_Pa.

Then we analyzed the correlation values and listed the features which have the strongest correlation with the target as well the stronges correlation pairs among the features:

```
## Strongest correlation between the target:
## NumHBondDonors           0.689196
## NumOfConf                0.513653
## hydroperoxide            0.314053
## hydroxyl (alkyl)         0.310452
## NumOfAtoms               0.307337
## carboxylic acid          0.304259
## NumOfC                   0.262769
## carbonylperoxynitrate    0.223739
## MW                       0.199574
## NumOfN                   0.183152
## Name: log_pSat_Pa, dtype: float64
```

```
## ---------------------
```

```
## Strongest correlation pairs:
## NumOfO                 MW                0.880358
## NumOfC                 NumOfAtoms        0.838402
## NumOfN                 MW                0.772575
## NumOfAtoms             MW                0.707009
## nitrate                NumOfN            0.687224
## NumOfN                 NumOfO            0.656750
## hydroxyl (alkyl)       NumHBondDonors    0.632023
## carbonylperoxynitrate  NumOfO            0.510409
## nitrate                NumOfAtoms        0.492108
## NumOfN                 NumOfAtoms        0.491902
## dtype: float64
```



Trying out different models

Baseline values without any feature engineering or other tweaking:

```
##                        Model  Train Loss  CV Loss Mean  Train R^2    CV R^2
## 0             DummyRegressor    9.735229      9.736501   0.000000 -0.000364
## 1           LinearRegression    2.892566      2.898423   0.702876  0.702194
## 2      RandomForestRegressor    0.388773      2.773706   0.960065  0.715013
## 3  GradientBoostingRegressor    2.611631      2.679421   0.731734  0.724689
```

Dummy

Linear Regression

Random Forest

Gradient Boosting Regressor