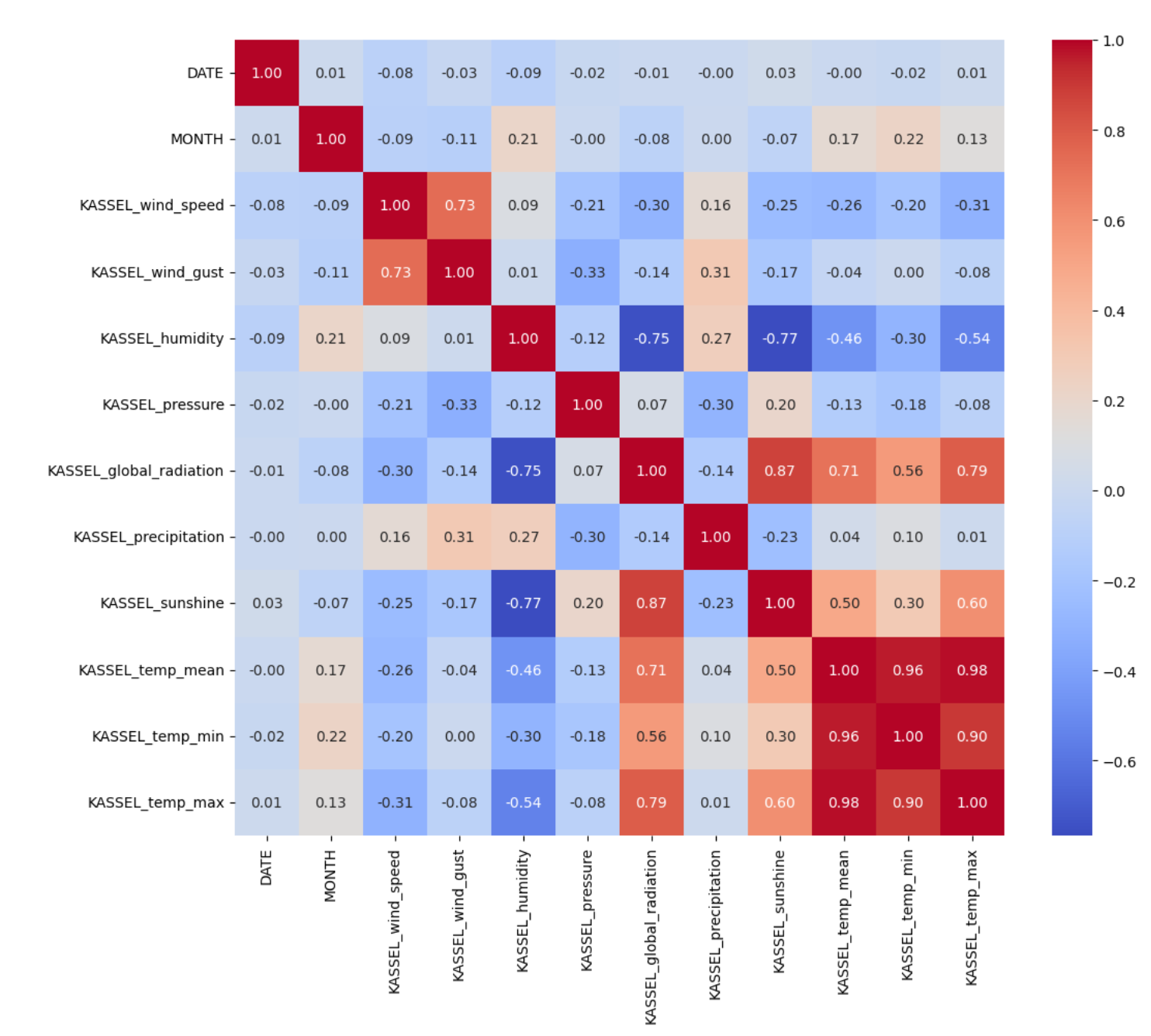
*Enhanced Weather Forecaster for Europe Final Report*

The current weather forecaster for Mainland Europe is out of date. As a result, it has lost accuracy over time and is unable to keep up with the new data that is being fed into the predictor.

The proposed solution was to scrap the current model entirely and come up with a completely new model. If the old model was both inaccurate and not learning from new data, it needed to be retired.

To set up the groundwork for conducting an analysis to create a better model, an existing weather dataset containing temperature values and weather attributes such as humidity, sunshine level, wind speed, etc. for every day across 18 different locations in Europe from 2000/01/01-2010/01/01 was used to create a training and testing dataset for the potential models.

To simplify the analysis, the model will only be predicting the mean temperature of tomorrow, or one day in advance. During the data wrangling and EDA phase, it was discovered that, to no surprise, the variables with the highest correlation to the next day’s mean temperature was the current day’s mean temperature alongside maximum and minimum temperatures. Other weather data that also showed some degree of correlation included global radiation and sunshine, as well as a negative correlation with humidity. The heatmap below provides greater clarity to the observed patterns.



The heatmap here depicts all of the variables for the location of Kassel, Germany, which is one of the 18 total locations available in the weather dataset. Note, not every location has the same set of variables, some are missing, but Kassel shown above is one where all weather variables were present.

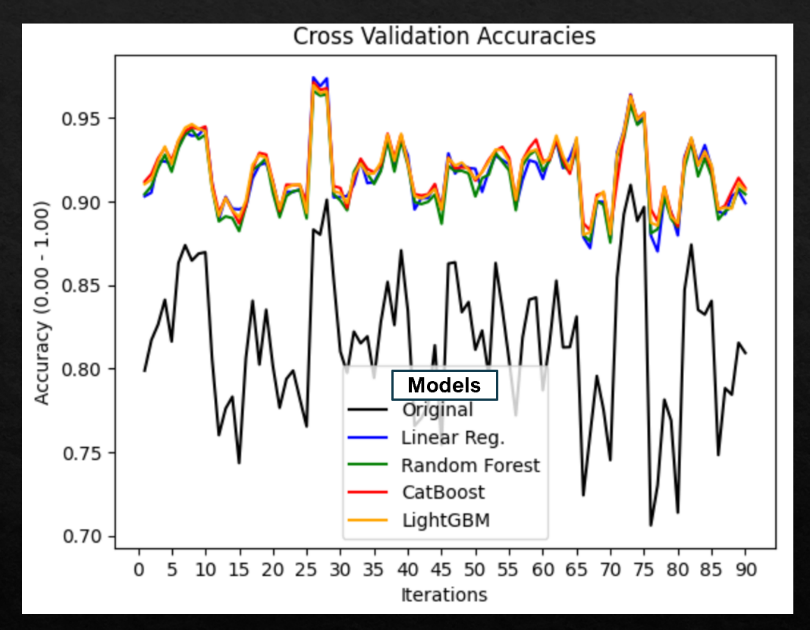
Afterwards, the dataset was broken down and rebuilt so that instead of each weather variable having an individual column by city, a new city column was created so that those weather variable columns could be merged for a more vertical dataframe.

The training and testing datasets were done in the form of a time series split, with 2000/01/01-2007/12/31 part of the training set and 2008/01/01-2009/12/31 the testing set.

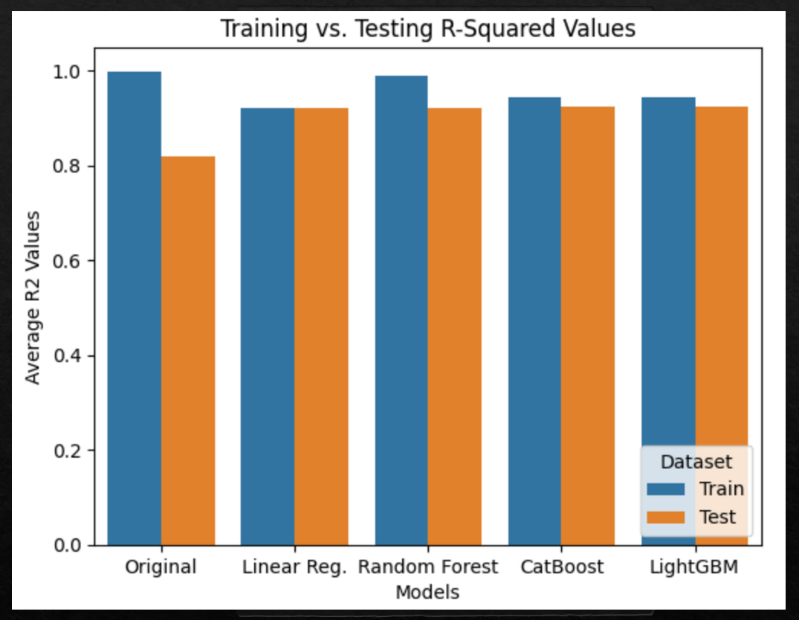
During the modeling phase, multiple models were experimented on and tuned to best optimize their accuracies and performance times. The original model was that of a decision tree. The models that were tested were as follows:

* Linear Regression
* Random Forest
* XGBoost (Extreme Gradient Boosting)
* CatBoost (Categorical Boosting)
* LightGBM (Light Gradient Boosting Machine)

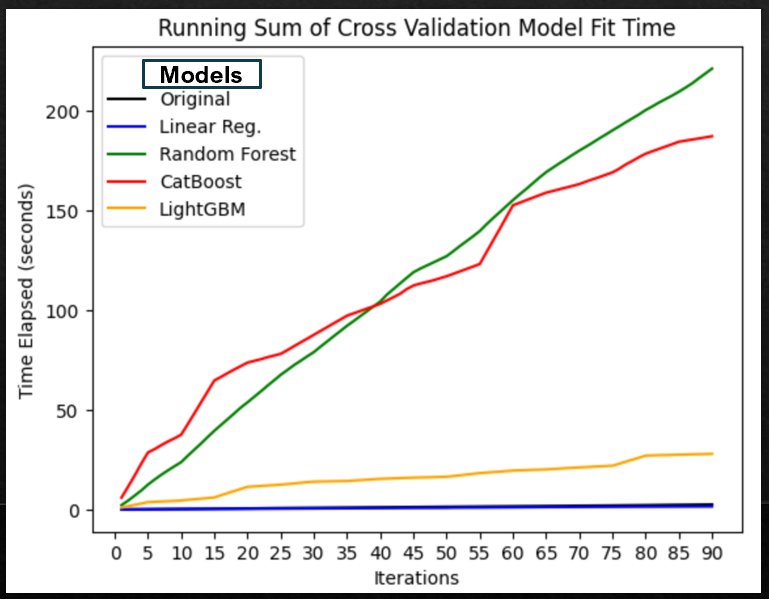
The conclusion that was reached was to select **CatBoost as the primary successor to the old model** and to have **LightGBM as the backup model**. Below are some figures that depict the accuracy and runtime of the models in question (XGBoost was omitted as it underperformed drastically on accuracy to be considered a replacement).



The figure shows the accuracy of each model across 90 cross validation iterations. All the tested models shown performed better than the original model, with CatBoost and LightGBM edging out the other two marginally.



The bar graph compares the average R-squared values between the training and testing set for each model. The original and random forest models show significant overfitting while the others have that problem much more controlled.



The line plot above shows the running total time taken for the model to predict after each iteration. As the iteration count goes up, the time disparity becomes increasingly obvious. Though CatBoost is accurate, it shows that against very large datasets it does not scale well.

The reason for passing on the linear regression model is that although both fast and accurate, it assumes a linear relationship between the weather patterns, which is an oversimplification and not always true. The random forest model can be seen to be the most computationally expensive.

The choice for CatBoost to be the primarly model comes down to the fact it is very easy to handle right from the start. There is little need for hyperparameter tuning and the model has built in features to prevent significant overfitting.

However, if future datasets the model will need to work with are much greater in quantity, both by location and weather variables, then switching to the LightGBM model will be the better option. Although not as easy to handle, LightGBM scales much better than CatBoost with large datasets. In terms of overfitting, LightGBM is only vulnerable on smaller datasets, so this would also no longer be a problem.

The decision to not use time series models such as seasonal autoregressive integrated moving average (SARIMA) was the fact that only one day in advance was being predicted. If the model was designed to predict further into the future, such as by 14 days, then such time series models could become more of use.

Also not tested in this analysis were models that involve neural networks. Likewise, with LightGBM, neural networks are much more effective predictors when the dataset is very large. Thus, if there is a further analysis to be made, time series or neural network models can be potential alternatives.

Conclusion of this analysis is to begin with CatBoost while data is limited. As it accumulates over time, switch to LightGBM. Each model has an average R-squared exceeding 0.9.