Results of Predicting and Forecasting Household Temperature

Name: Qinyun (Peter) Yu, 501137007

Date: August 4th, 2022

Supervised by: Dr. Alan Fung

Results

Background

Environment Canada provides weather data to varying degrees of granularity ranging from a timeline of 1840 to 2022. This data includes weather labels, dew point, temperature, wind speed, and other related features like humidity. This data comes from weather stations, various governmental buildings, and other buildings of significance like airports. By coupling weather data from several weather stations, we can potentially train machine learning models to help predict and forecast household dry bulb temperature. These developed models would then have several applications like being able to identify potential temperature sensor failures and incorporate predicted/forecasted household dry-bulb temperature into temperature-driven applications like HVAC systems. Which then leads to the greater picture of the work done, where in the future, forecasted temperatures can be used to aid in HVAC decision making and lead to a smarter thermostat system that does not just rely solely on a singular sensor, and one that relies on a few various external features like local weather station data.

As a statement for this report, because the discussion page is not uploaded, a part of the discussion for the MRP will be discussed.

Toronto Results – Comparison of Models

By using Toronto City as a “proxy” house, and use Toronto City Centre, Toronto International Airport, and Buttonville Airport weather data as exogenous features, we test 30 1-hour-lagged components on various models (those of which are specified in the methodology report, and the most significant features are later highlighted in the LightGBM section). The station names in the report are listed as is and are the official names on the Environment Canada site.

When looking at the SARIMAX results finetuned with auto-ARIMA, we see that a p, d, q value of 4, 1, 5 best represents our dataset. This auto-ARIMA-coupled training took 62 minutes to train. The other two models used, Prophet, and LightGBM both were trained near instantaneously, which is of something to note, as there are ~40,000 data points along with 30 features. The LSTM model was not able to be configured properly, likely due to difficulties with the input shape and TensorFlow configuration problems. Additionally, Facebook Prophet on a Windows 10 environment had several installation problems, and initial results were run on a MacOS environment.

When looking at the error metrics (RMSE, MAE, R^2), we see that the Facebook Prophet model outperformed both the SARIMAX and LightGBM model (Table 1). One thing to note however is that these are the initial results, and further hyperparameter tuning is necessary for the LightGBM model specifically as decision tree models tend to overfit.

Model	RMSE	MAE	R^2
SARIMAX	2.25	1.77	0.95
Prophet	0.70	0.50	1.00
LightGBM	0.81	0.58	1.00

Table 1: Results of using a 80:20 train-test dataset of data ranging from 2016-01-01 to 2022-05-01

Additionally, when we look at the resulting plot, having overlayed the various models on each other, we see that for the most part, it seems ARIMAX is under-forecasting, while models like LightGBM and Prophet seem to generalize the data better (Figure 1).

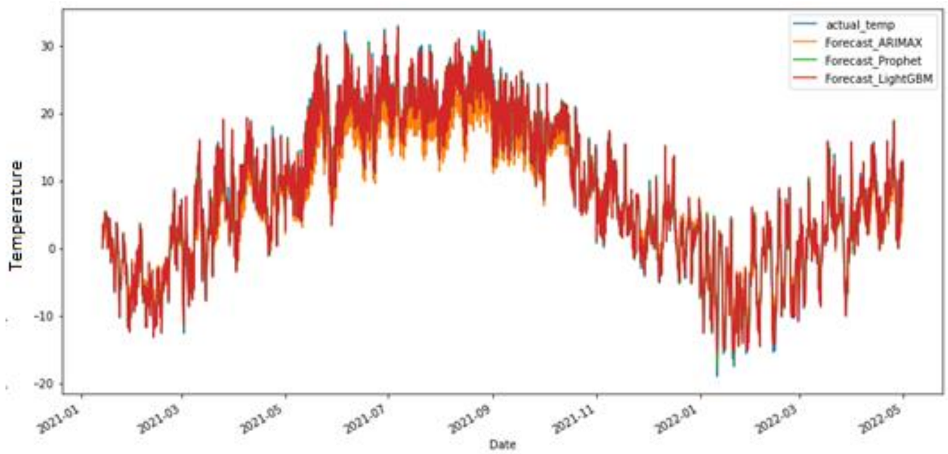

Figure 1: Model forecasts of temperature data from test data from 2021-2022

Toronto Results – Comparison of Time-of-Day Forecasting and Seasonal Forecasting

Feature Importance

Due to the inherent nature of decision tree models (LightGBM) of being able to improve results potentially drastically with hyperparameter tuning, being able to observe feature importance, lower memory usage, and high scalability, the following model results were generated in LightGBM.

When looking at the feature importance of the summer daytime model we see that incorporating solar radiation to our model is more important than using the 1-hour lagged temperature, and that geospatial features like local weather station temperature, wind speed, and dew point greatly influence our model (Figure 2a).

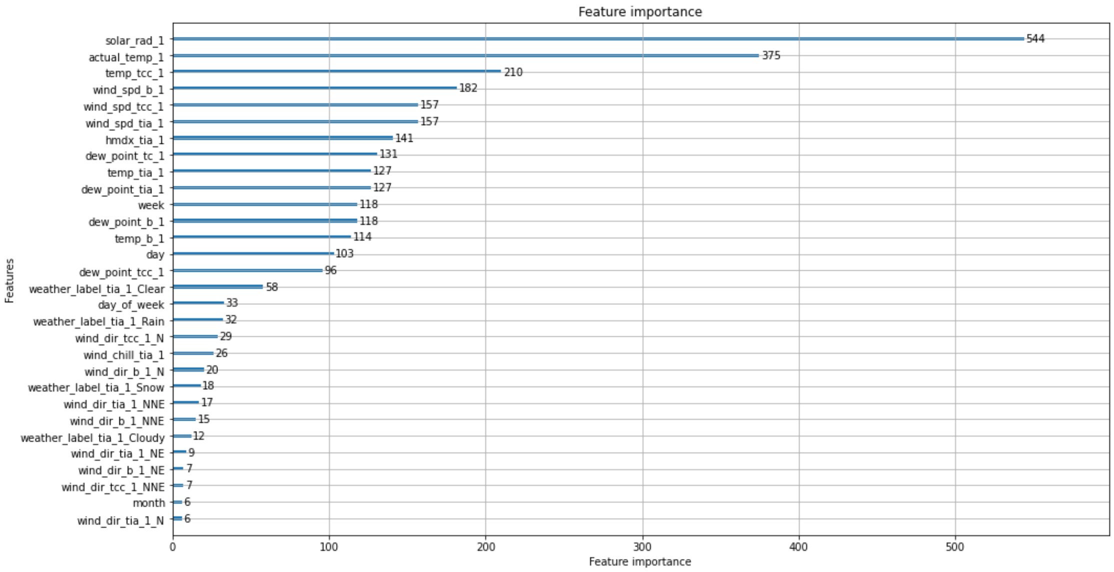

Figure 2a: LGBM feature importance of Toronto summer day model with lagged solar radiation data.

Accordingly, when looking at the following feature importance plots, we see repeatedly, that features like solar radiation (during the day), ground truth temperatures, day of the week, and dew point temperature are the greatest factors in our models (Figure 2b-d).

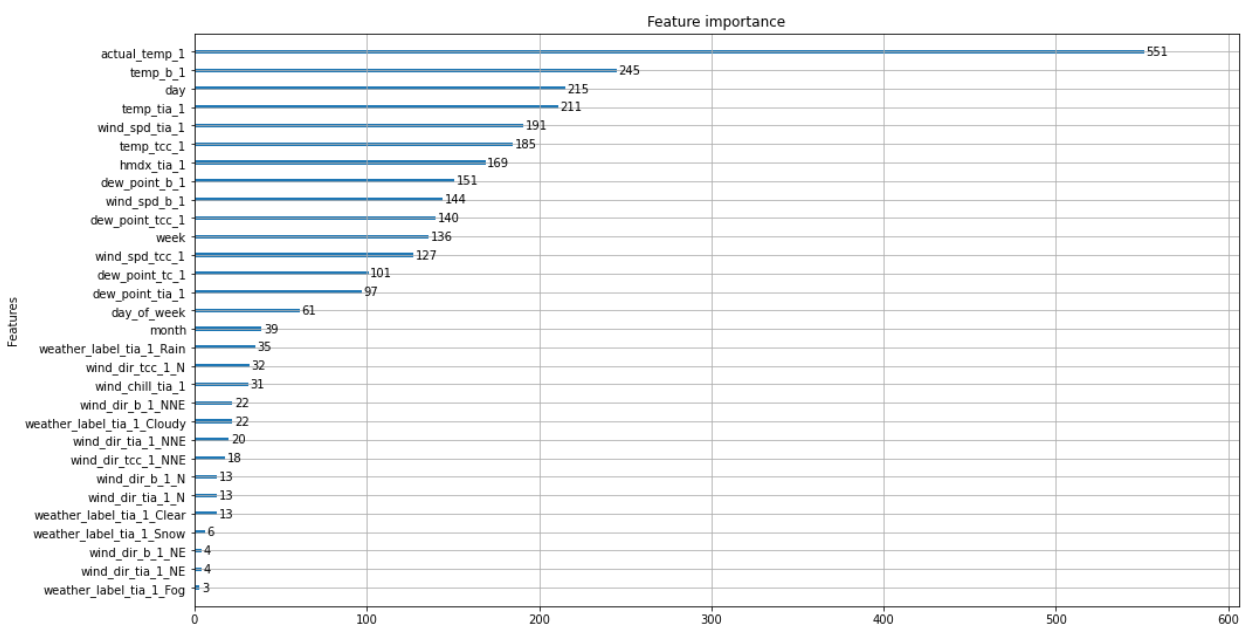

Figure 2b: LGBM feature importance of Toronto summer night model

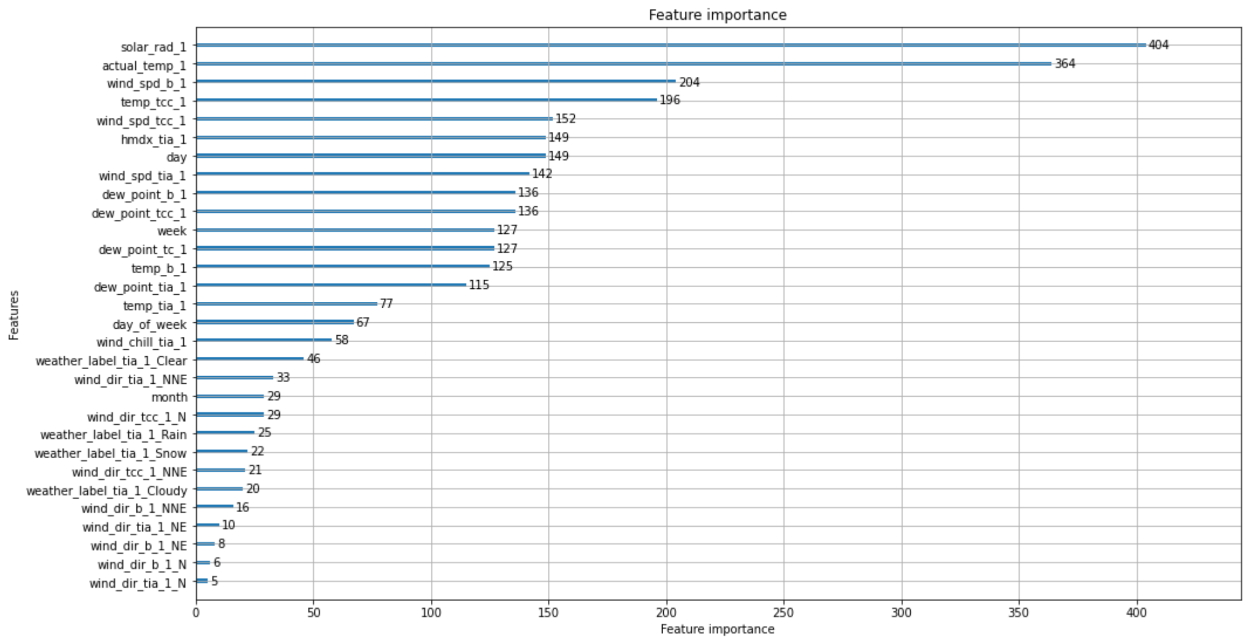

Figure 2c: LGBM feature importance of Toronto winter day model

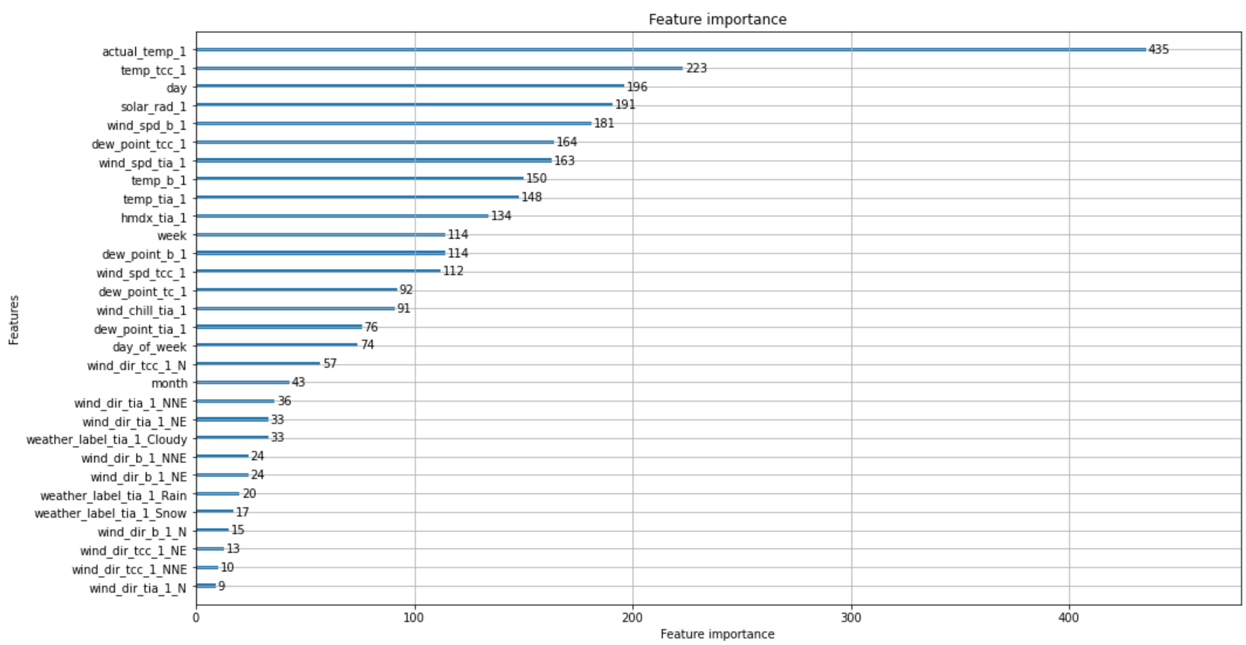

Figure 2d: LGBM feature importance of Toronto winter night model

King City North Results

When using the same three weather stations as exogenous features (Toronto City Centre, Buttonville Airport, Toronto International Airport), and changing our "proxy" household from Toronto City to King City North, we see that the same features are prioritized in our LightGBM models. The following feature importance plots will be provided in the supplementary material section.

Vancouver Results

When retraining our LightGBM model on a similar setup in Vancouver (i.e., Vancouver Harbour CS as the "proxy" household, West Vancouver Aut, Vancouver Intl Airport, and Point Atkinson as the exogenous stations used for features), we see again that the same features are emphasized in the LightGBM model, namely the geospatial features, along with temporal features like time of day, and time of week. These will also be provided in the supplementary section.

Error Metric Results

When looking across the board on model performance, we see that when handling 1-hour lagged forecasts, the model is decently accurate when given nearby station data (Table 2). Having moved our household ~25 km further from Toronto City to King City North results in decreased accuracy, and that in the future, for improved model results, one needs to obtain relatively close weather station data for more accurate predictions and forecasts (the feature importance data is found in the supplementary section).

	Summer + Day	Summer + Night	Winter + Day	Winter + Night
RMSE Toronto	0.88	0.61	0.66	0.60
RMSE King City North	1.07	0.75	0.84	0.78
RMSE Vancouver	0.80	0.65	0.65	0.57
MAE Toronto	0.65	0.42	0.49	0.38
MAE King City North	0.80	0.51	0.62	0.50
MAE Vancouver	0.59	0.47	0.46	0.39
R2 Toronto	0.99	0.99	1.00	1.00
R2 King City North	0.99	0.99	0.99	0.99
R2 Vancouver	0.98	0.98	0.98	0.98

Table 2: Error metrics of LightGBM model across various locations in Canada

Author's Note

At the time of August 1st, 2022, the RETScreen NASA data was not available so future studies and research conducted may not have solar radiation data available. Additionally, further studies are currently undergoing regarding hyperparameter tuning, decision tree feature selection, testing various lag periods with model results, and dataset size with model accuracy.

Supplementary Information

King City North Feature Importance

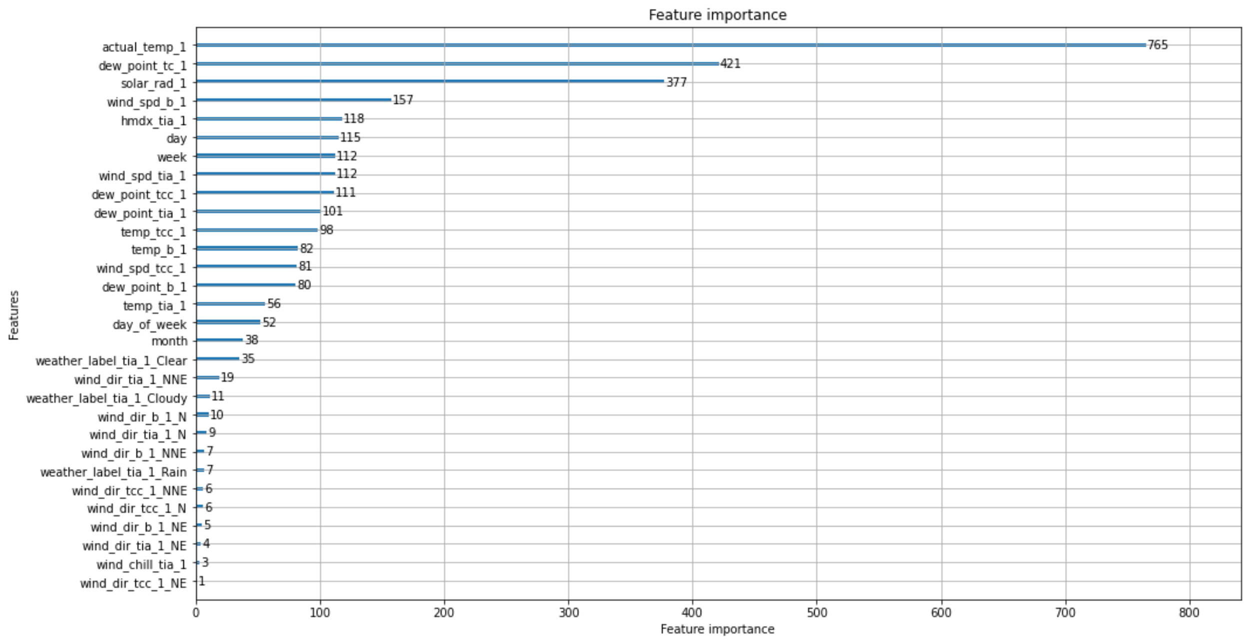

Figure 3a: LGBM feature importance of King City North summer day model

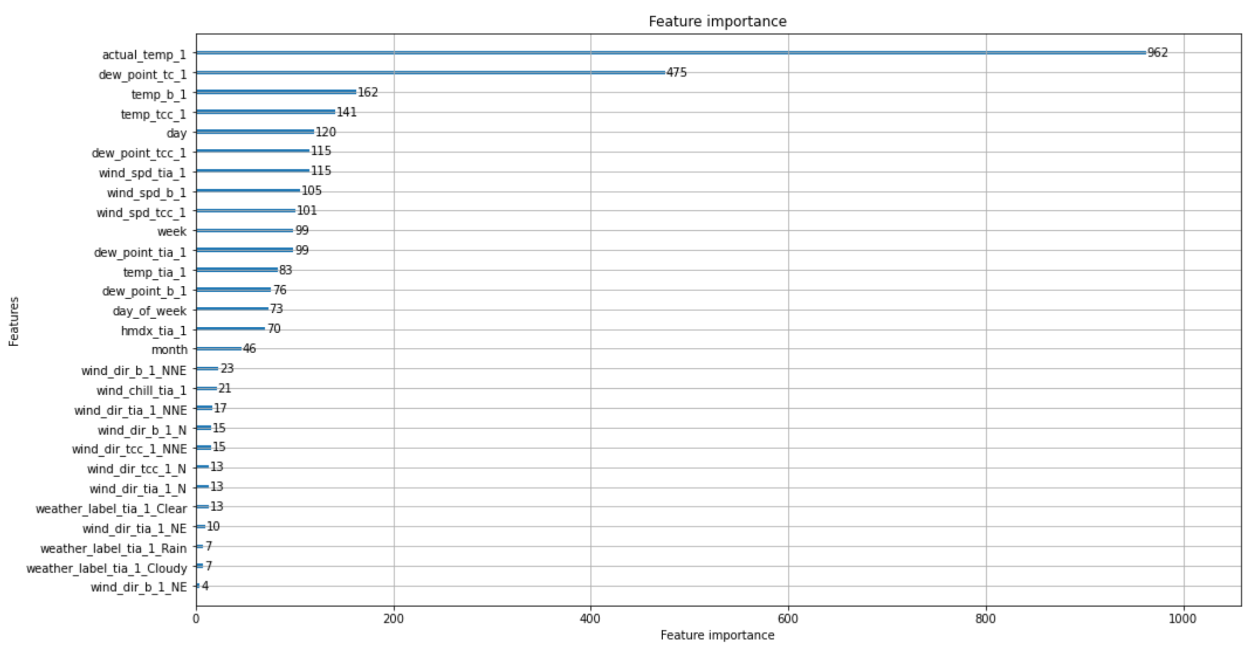

Figure 3b: LGBM feature importance of King City North summer night model

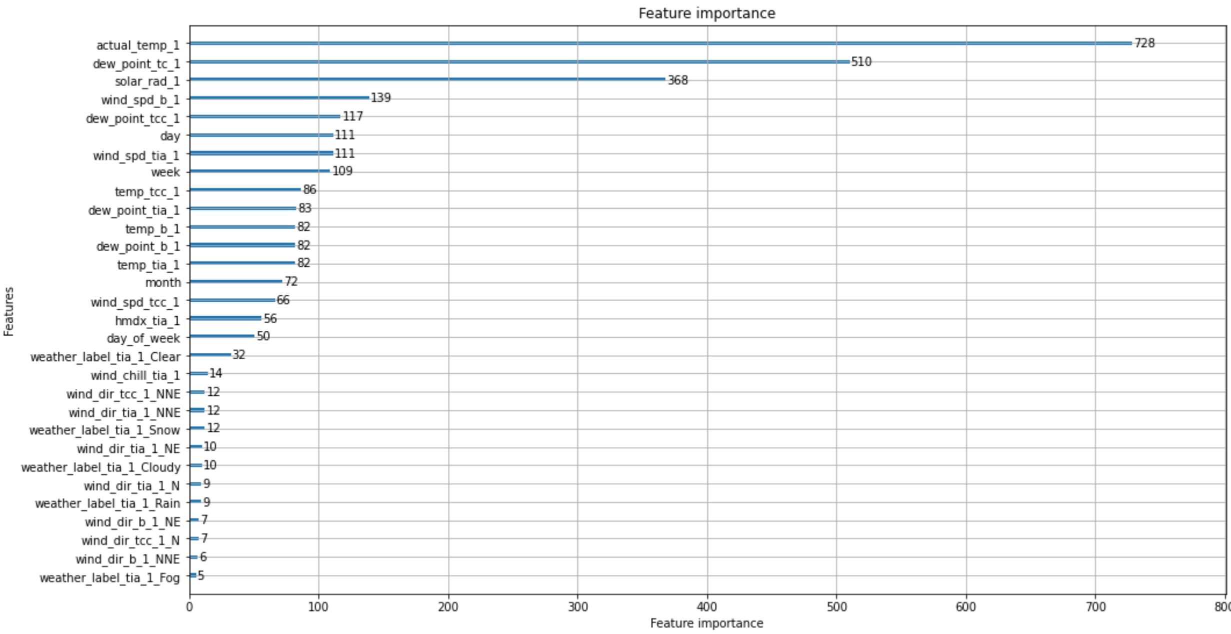

Figure 3c: LGBM feature importance of King City North winter day model

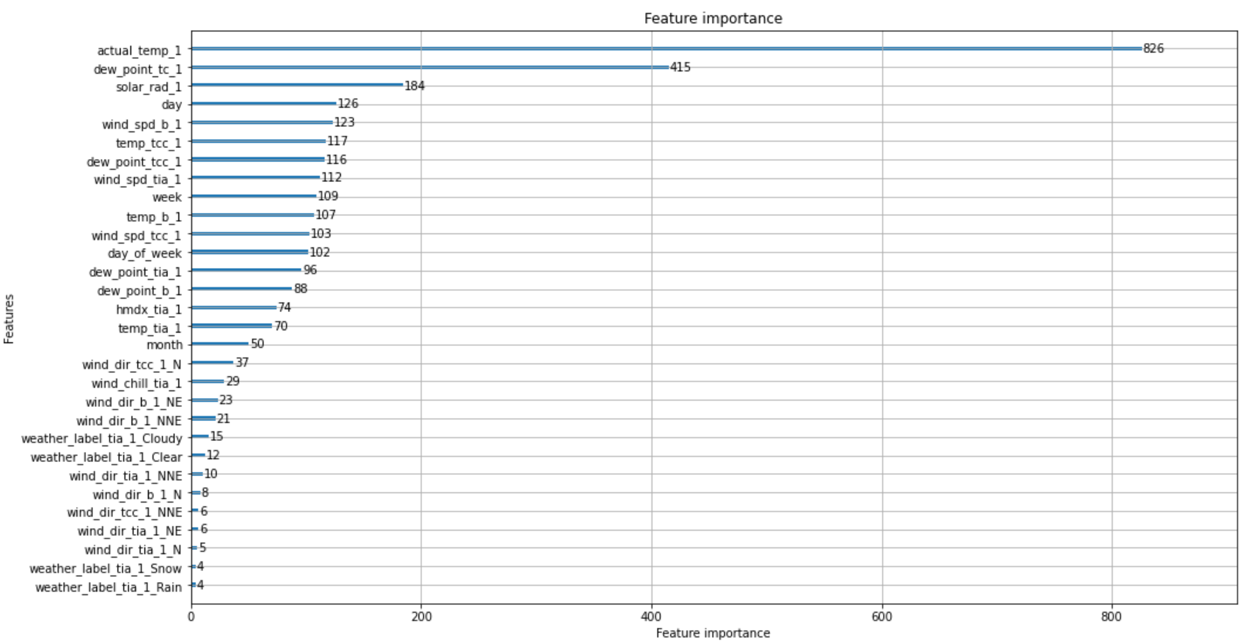

Figure 3d: LGBM feature importance of King City North winter night model

Vancouver Feature Importance

North Feature Importance

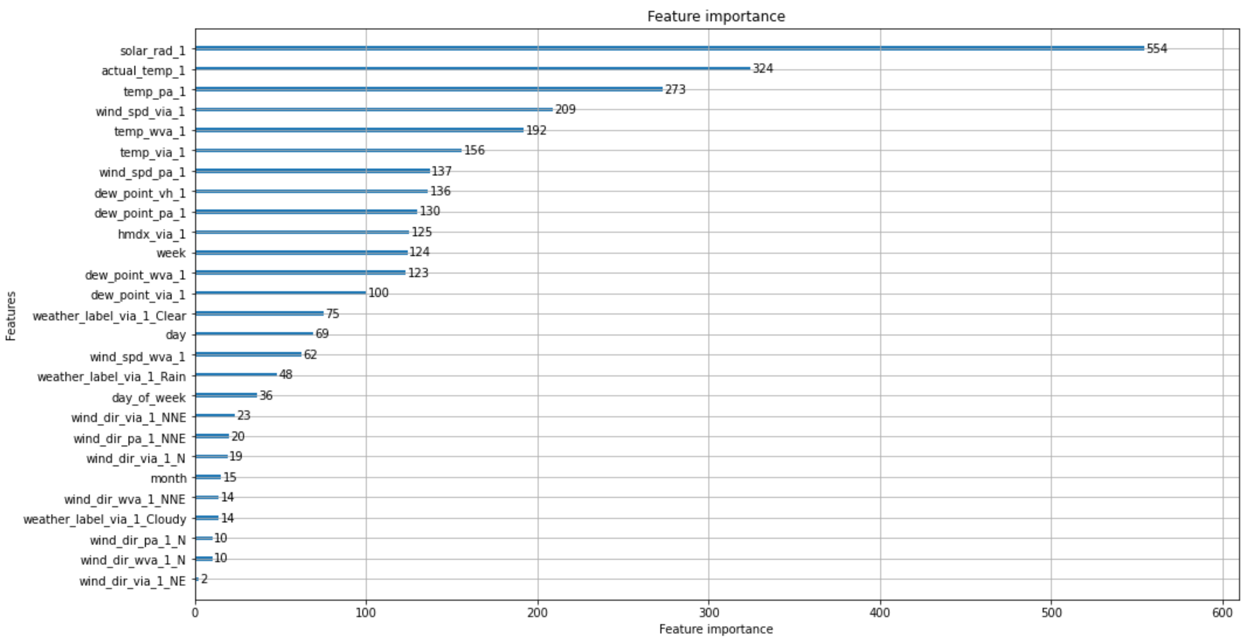

Figure 4a: LGBM feature importance of Vancouver summer day model

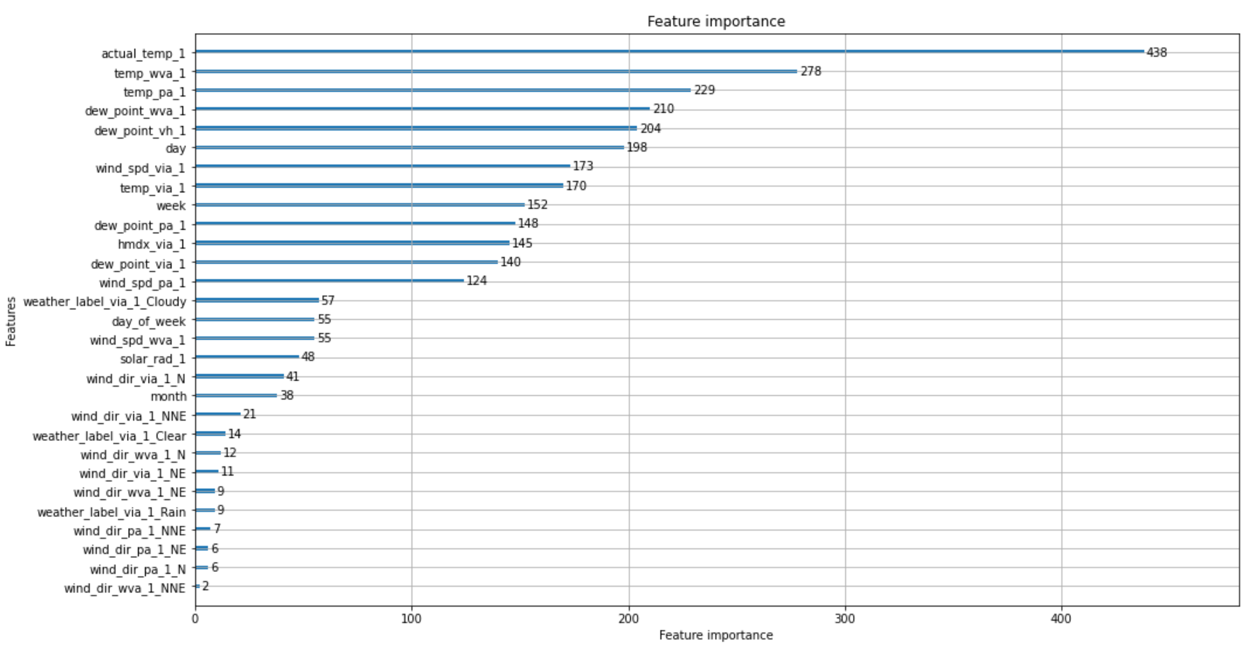

Figure 3b: LGBM feature importance of Vancouver summer night model

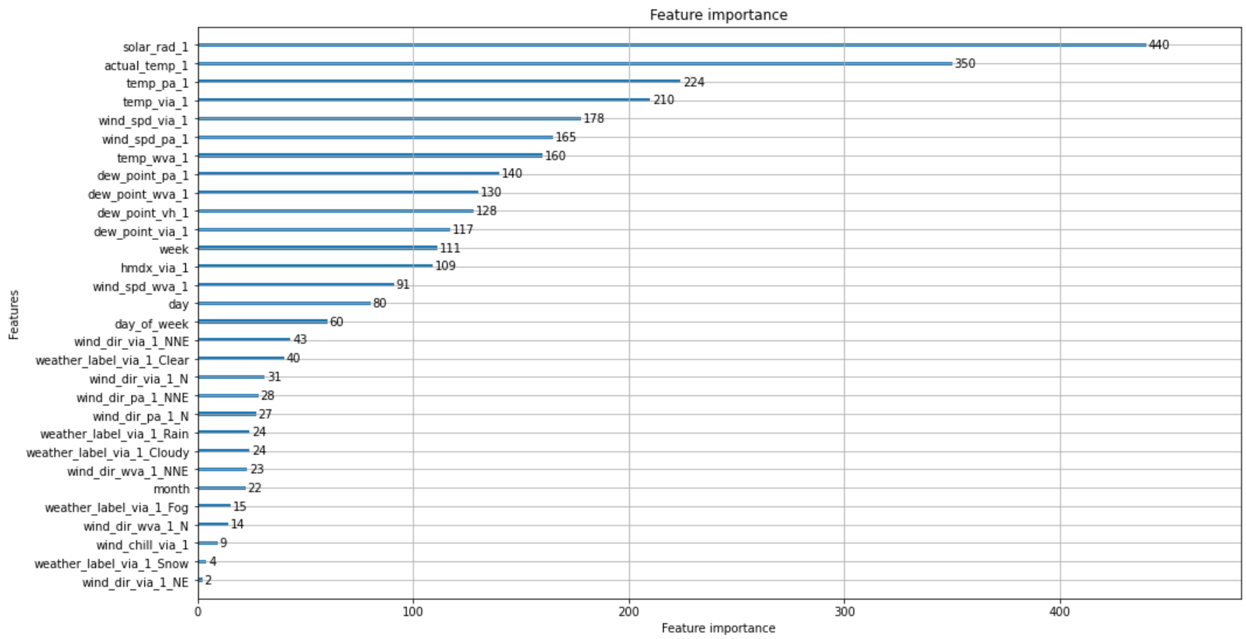

Figure 3c: LGBM feature importance of Vancouver winter day model

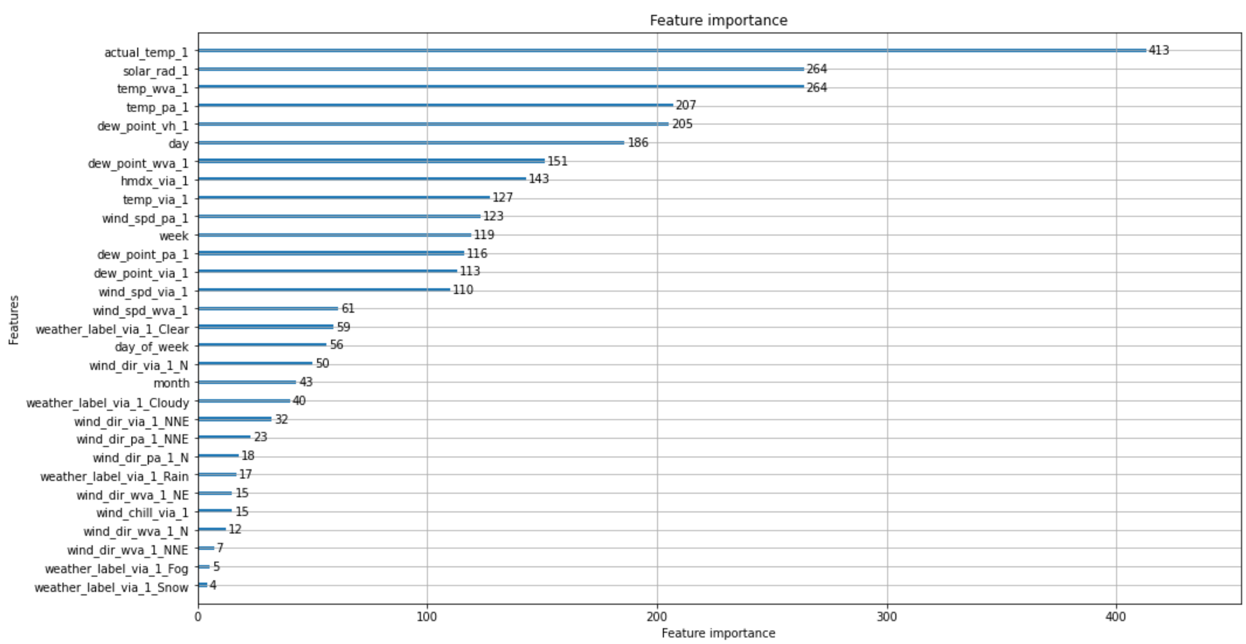

Figure 3d: LGBM feature importance of Vancouver winter night model