For the Feburary 2024 forecast, I updated the methodology to reduce volatility of the forecast from month to month,reduce reliance on unreliable predictor variables, and to increase the intuitiveness of forecasted values

The changes are annotated in red within the text below. However, the main changes impacted the following facets of the forecast.

1. I eliminated unreliable predictor variables, specifically: max of maximum temperature, min of minimum temperature, maximum snowfall, days of rain, and days of snowfall

2. Instead of deseasonalizing individual place series and then one period differencing them, I take a 12 month seasonal differencing. The predictor variables are also differenced by 12 months. Therefore, changes from the same month last year are due to changes in weather and gasoline prices from last year

3. Instead of regressing the difference of the aggregate series on the changes of the individual place series to get aggregate forecasts, I regress on the levels instead after using the place changes to derive place level values. I also constrain the impact of each place level value on the aggregate series to be positive. This is necessary due to variance from pull to pull in google trends causing some places to have a negative impact on aggregate search interest

4. I normalize the aggregate series such that the historical minimum value is zero and the historical maximum value is 100 instead of the previous technique that only normalized the maximum value. This has no practical impact but aids in interpretation.

Northern Michigan Search Interest: Forecast Implementation

==

Over the past eight installments, I have pulled 2017-2023 search interest data for 141 Lower Peninsula and 71 Upper Pensinsula cities and towns from Google Trends using the pytrends python package. I have performed a myriad of time series and cross-sectional analysis on these data series. I found that seasonal variation explains most changes in search interest for the places. For a majority of places, the highest interest occurs in the summer while many Upper Pensinsula places are have a secondary high interest point in the winter probably due to snowmobiling. I also found that unseasonable weather and gas prices explain residual search interest variation for some places. Finally, I found that some places are growing in search interest. This is especially true for places in the relatively low population density eastern portion of the northern Lower Pensinsula and the relatively middle population density areas of the Upper Pensinsula along Lake Michigan and the Wisconsin border.

It's my goal to provide three months search interest forecasts on a monthly basis. These will be provided for each town and city and combined for each peninsula. In order to do this, I must collect current values for search interest for each place and create three month forecasts for each of the weather and gasoline price variables. Then, the machine learning lasso models will be recalibrated on the new data and used to predict future search interest.

Update Feb 2024: I now use data from 2016 to use in 12 month differencing so I don't lose a year of data.

==

Updated Data

First, updated data for each of the 212 towns and cities is pulled up to the most recent month from Google Trends using the python package pytrends. These data series are pulled twice. The first time each place's data series is pulled individually such that each data series is normalized from 0 to 100, where 100 is the maximum value in the total series. These are the data series used to model each individual place. The second time data is pulled such that all data series are on the same scale (i.e. each peninsula's most popular place's search interest maximum is 100). This is done per the article: https://medium.com/analytics-vidhya/compare-more-than-5-keywords-in-google-trends-search-using-pytrends-3462d6b5ad62. These second data series are then combined by peninsula and normalized such that the maximum of the entire series is 100. This is used to model combined peninsula trends.

After the search interest data is updated, the following transformations are performed:

1. The data is deseasonalized
2. The data is differenced one period to account for trend stationarity
3. The Covid-19 period (March-August 2020) is eliminated

Weather data for the most recent month is pulled from weather.gov. For the Lower Peninsula, weather data is sourced from Houghton Lake. For the Upper Peninsula, weather data is sourced from Sault Saint Marie. These locations were chosen because they have average temperature data in weather.gov's most easily downloadable format, instead of just min and max temps. Finally, the latest gasoline prices are downloaded from St. Louis Fed website.

As described in the previous installments, there are several transformations performed on all of the predictor variables. For starters, it's already clear that average seasonal temperature drives search interest as the search interest data is highly seasonal. However, by the time that I predict search interest with weather, it is already deseasonalized. Therefore, what we are really interested in is unseasonable weather patters. Therefore, the monthly mean of each weather variable is subtracted from the values.

Since the model is differenced, all of the weather and gasoline variables are differenced. Since a lasso model will be employed for place level predictions, the predictor variables are finally standardized by mean and standard deviation.

Update Feb 2024: Now the minimum value of the peninsula aggregated series is normalized to zero instead of only normalizing the maximum value to 100.

Update Feb 2024: Aggregated data is no longer differenced or deseasonalized in any way. It is aggregated based on the level actuals and forecasts of the individual places.

Update Feb 2024: Place level data series are no longer deseasonalized and 1 period differenced. Instead they are seasonally differenced by 12 months. For 2021 this is 24 months because 2020 is eliminated

This is for at least two reasons

1. Due to the seasonal nature of northerm Michigan search interest, the last actual value often has tentative relationship with the three month forecast. For instance, a low April forecast will not necessarily reflect in low May, June, July search interest because summer months have different search interest dynamics than April

2. Deasonalizing and differencing the data series led to excess volatility. This is normally done to address trend stationarity issues which is found especially in the Upper Peninsula series. However, in practice I decided this concern was less important.

Update Feb 2024: The monthly mean of each weather variable is no longer subtracted from each variable. They are also not 1 period differenced. Instead, they are 12 month or seasonally differenced. The values are still standardized by mean and standard deviation as is generally required in regularized regression.

==

Predictor Forecasts

Since we are forecasting search interest based on weather and gas prices, it is necessary to have three month forecasts for these factors. Naturally, this requires a fair amount of subject matter knowledge. Therefore, I either get these forecasts from reliable sources, or abstract from them.

Minimum and Maximum Temperatures. Days of Precipitation and Snowfall.

I source these variables from accuweather.com monthly forecasts. These are the variables/transformations sourced from accuweather:

1. Maximum temperature: Monthly average and maximum
2. Minimum temperature: Monthly average and minimum
3. Days of precipitation
4. Days of snowfall

Update Feb 2024: I now longer use max of maximum or min of minimum temperatures. Using historical data, these variables are often good predictors probably because extreme temperatures are news worthy. But my method of producing these forecasts never matched the true extremes of temeratures. Since the accuweather forecasts are daily expectations, it is obvious that the maximum of the daily expectations will not equal the actual maximum expected to be achieved during the month. The same can be said for minimums.

Update Feb 2024: I no longer use days of rain or snowfall. These always produced high values especially for the third month out. I suspect that the probability threshold used by accuweather to predict a rain or snow day must be very low.

Average Temperature and Average Precipitation

I abstract these values from weather.gov climate forecasts. Weather.gov provides the probabilities that temperature and precipitation will be above or below established 30 year normals. The monthly normals for temperature and precipitation for each location are also provided by weather.gov.

To transform weather.gov normals and probabilities to temperature and precipitation values, I also take the standard deviation for each of these values per calendar month based on the available 20 years from the same period the normals are calculated under. (Note the normals are calculated from 1990-2020, so I use 2000-2020 to calculate the standard deviations). Note that precipitation values have a higher standard deviation in the summer and temperature values have a higher standard deviation value in the winter.

Finally, I assume that temperatures and precipitation follow a normal distribution. Assuming this distribution, using the established normals, and using the standard deviations and probabilities that temperatures or precipitation is above or below normal, it is possible to use an inverse normal cumulative distribution function (CDF) to calculate predictions for each of these variables.

Using these assumptions, the higher the probability that a value is above or below established normals and the higher the historical standard deviation for that month will lead to forecasts that differ more from the established normals.

See equations below. Here mu is the historical normal for month 'm' and variable 'k' and sigma is the standard deviation. P is the probability value where P<0.5 would be below the historical normal and P>0.5 would be above. Here the mean of the normal distribution is 0.

$$\mu_{mk} + N^{-1}(P_{mk}, 0, \sigma_{mk})$$

Average Snowfall and Snow Depth

Snowfall and snow depth are not directly provided by accuweather or the weather.gov climate forecasts. Instead, I use a linear regression of each variable on the month, average minimum temperature, average maximum temperature, average precipitation, and days of snowfall. These models are estimated on historical data from weather.gov and forecasted based on forecasts for these four weather variables created as per above from accuweather or weather.gov climate predictions.

I also restrict the regression to only months October through April. All other months are assumed to be 0 for all snow related variables. The following variables are forecasted in this way:

1. Average snowfall
2. Average snow depth
3. Maximum snow depth
4. Days with any snow depth
5. Days with at least six inches of snow depth

The following equation represents that used to estimate each of these snow related variables.

Update Feb 2024: In my original writeup, I failed to mention the maximum snowfall variable that I also used a linear regression to calculate. I am removing this variable now. It had a very low r-squared (around 0.3). It's difficult to predict what that highest single day of snowfall will be in a month.

$$SNOW_t = \beta_0 + \beta_1 * MINTEMP_t + \beta_2 * MAXTEMP_t + \beta_3 * PRECIPAVG_t + \beta_4 * SNOWDAYS_t + \beta_5 * MONTH_t + \varepsilon_t$$

As per the table below, the regressions are fairly successful at predicting the snow variables with in sample r-squared ranging from 0.54 to 0.90. The regressions predict Houghton Lake days of snow depth over six inches the worst. However, for Sault Ste. Marie this variable is predicted much more effectively.

	City	Variable	R-Squared
0	Houghton Lake	Average Snowfall	0.71
1	Houghton Lake	Average Snow Depth	0.74
2	Houghton Lake	Max Snow Depth	0.74
3	Houghton Lake	Days Any Snow Depth	0.90
4	Houghton Lake	Days 6 Inches Snow Depth	0.54
5	Sault St. Marie	Average Snowfall	0.78
6	Sault St. Marie	Average Snow Depth	0.85
7	Sault St. Marie	Max Snow Depth	0.75
8	Sault St. Marie	Days Any Snow Depth	0.90
9	Sault St. Marie	Days 6 Inches Snow Depth	0.90

Gasoline Prices

Gasoline prices are forecasted using an ARIMA model. In order to make the gasoline price series stationary I deseasonalize the data, take the natural log of the values, and then difference the series. After taking these steps, the series follows a MA1 process with a positive moving average coefficient. In other words, values in the previous period higher or lower than what the model predicted will be repeated in the current month. After making predictions using this model, I reverse the differencing, natural log transformation, and deseasonalization.

$$\Delta ln(Gasoline_t) = -0.01 + 0.55 * \varepsilon_{t-1} + \varepsilon_t$$

Predictor Variable Transformations

Predictor variables are transformed just as the actual values as described above. However, the means and standard deviations that are used for transformations are acquired from the actual values.

===

Model Recalibration

As discussed in previous installments, each of the place series is modeled using a lasso regression segmented between warm weather months (May-October) and cold weather months (November-April). In other words, a separate model is estimated for each place and each group of months (warm/cold). The model parameters are re-estimated and also the hyperparameters are re-estimated using the cross-validation techniques discussed in previous installments.

Once the place models are re-estimated, the forecasted predictor variables created (as described above) are used to create three month forecasts for each place. These place level forecasts are then un-differenced based on last actuals and re-seasonalized to create forecasts for each individual place.

However, I also create aggregated forecasts for each peninsula. To do this, the original differenced, deseasonalized actual values for each place are normalized by mean and standard deviation. Then ridge regressions are estimated to predict the aggregated pensinsula series based on the individual places. Only the model parameters are re-estimated (as the hyperparameters are generally very small).

Once the ridge regressions are estimated, the predicted place level values are also standardized by mean and standard deviations and used to make peninsula level predictions.

Update Feb 2024: As place level forecasts are no longer de-seasonalized, re-seasonaization is no longer necessary. Un-differencing or unwinding is now based on the difference from the value of the same month the previous year as opposed to the previous month

Update Feb 2024: To create the aggregated forecast, I now regress the levels of aggregate search interest on the place search interest levels not the differences. I get the place interest levels by unwinding the forecast based on last year's actuals and forecasted changes, as decribed above.

Pictured below is the lasso optimization problem for each place level regression with search interest y, predictor variables X, model parameters beta/theta and hyperparameters alpha.

$$\left\{\begin{array}{ll} May - October & \sum_i^n(\Delta y_i - (\beta_0 + \sum_j^k \beta_j \Delta X_{ij})) + \alpha_1 \sum_j^k |\beta_j| \\ November - April & \sum_l^m(\Delta y_l - (\theta_0 + \sum_j^k \theta_j \Delta X_{lj})) + \alpha_2 \sum_j^k |\theta_j| \end{array}\right\}$$

Update Feb 2024: This equation is largely accurate but the delta should be understood as 12 month not 1 month changes

Pictured below is the ridge optimization problem for each peninsula with search interest U/L, predictor variables y, model parameters p/v, and hyperparameters alpha.

$$\left\{\begin{array}{ll} Upper Peninsula & \sum_o^p(\Delta U_o - (\rho_0 + \sum_t^u \rho_t \Delta \widehat{y}_{ot})) + \alpha_3 \sum_t^u \rho_t^2 \\ Lower Peninsula & \sum_q^r(\Delta L_q - (\upsilon_0 + \sum_v^w \upsilon_v \Delta \widehat{y}_{qv})) + \alpha_4 \sum_v^w \upsilon_v^2 \end{array}\right\}$$

Update Feb 2024: As discussed in the notes above, there are no longer any deltas or changes in this equation. Aggregate search interest levels are regressed on place search interest levels. Also, the estimated parameters are constrained to be positive. Previously, variation in Google trends data from data pull to data pull led to negative impacts of some individual places on the aggregate search interest.