Ontario Energy Use: Predicting demand from weather and population data

James Parks

1 Background

Ontarians use more than 140 million megawatt hours of electricity a year to run their homes, businesses, hospitals, schools, transit and infrastructure. It is clear that weather and population play a huge influence in the demand for energy. In order to meet client demands for electricity, energy companies need a robust model. More energy is typically demanded during the day than at night when people are sleeping, but since the change in weather can also affect changes in demand for energy, companies need to be able to anticipate large spikes in demand in order to prevent blackouts.

The Independent Electricity System Operator (IESO) operates and settles Ontario's electricity wholesale market, where the market price is set based on accepted offers to supply electricity against the forecasted demand. In this project we look at the historical hourly energy demand of Ontario, Canada from January 1994 until August 2017, the historical energy price, both from the IESO, the historical population as well as the historical hourly weather data for Toronto from the same time period from Environment Canada's historical climate data.

The goal of this project is to develop a model that predicts energy demand based on weather, price and population data.

1.1 The Client

The client in this case is the IESO, or related organizations in other regions with similar weather patterns and populations. Having a more robust model for determining just how much changes in weather affect energy demand can be extremely beneficial in saving money and preventing blackouts. Based on this analysis they will be better able to adjust supply to meet expected demands with the changes of seasonal weather and population growth.

2 The Data

The Data comes from three sources.

Independent Electricity System Operator, IESO power data which has .csv files for the hourly energy demand for the province of Ontario from January 1994 to August 2017, hourly price from 2002 to 2017 and hourly import and export data.

Environment Canada, which has .csv files for the hourly weather data for many locations in Ontario for the same time period. To start with we choose the data for Toronto, since its urban area accounts for a large percentage of the population of Ontario and we make the assumption that its energy demand will contribute to the largest demand in the province.

Wikipedia, which has the population data for Ontario and Toronto for the years 1996, 2001, 2006, 2011 and 2016. We will use linear interpolation to estimate the approximate population for the intermediate time periods.

3 Exploratory Analysis

From the Environment Canada weather dataset we have the following weather Variable Descriptions features.

3.1 Variable Descriptions

- **Date/Time**: Date and Time given in local Eastern Time.
- **Data Quality**: Is either ' ' if there is no problems with the data quality or nan if the weather data is missing.
- **Temp (°C)**: The hourly temperature give in Celsius
- **Temp Flag, Dew Point Temp Flag, Rel Hum Flag, Wind Dir Flag, Visibility Flag, Stn Press Flag, Hmdx Flag or Wind Chill Flag**: Is either 'M' or nan if the corresponding feature is missing.
- **Dew Point Temp (°C)**: is the temperature to which air must be cooled to become saturated with water vapor. A high relative humidity implies that the dew point is closer to the current air temperature.
- **Rel Hum (%)**: is the ratio of the partial pressure of water vapor to the equilibrium vapor pressure of water at a given temperature.
- **Wind Dir (10s deg)**: is the direction from which the wind originates. It is measured on a scale of 0 to 360° in 10 degree increments. Here North

is between 35. and 1.

- **Wind Spd (km/h)**: is the wind speed given in kilometers / hour.

- **Wind Spd Flag**: Is either 'M' or nan if the corresponding feature is missing or 'E' if the speed is given, but not the direction.

- **Visibility (km)**: is a measure of the distance at which an object or light can be clearly discerned, given in kilometers.

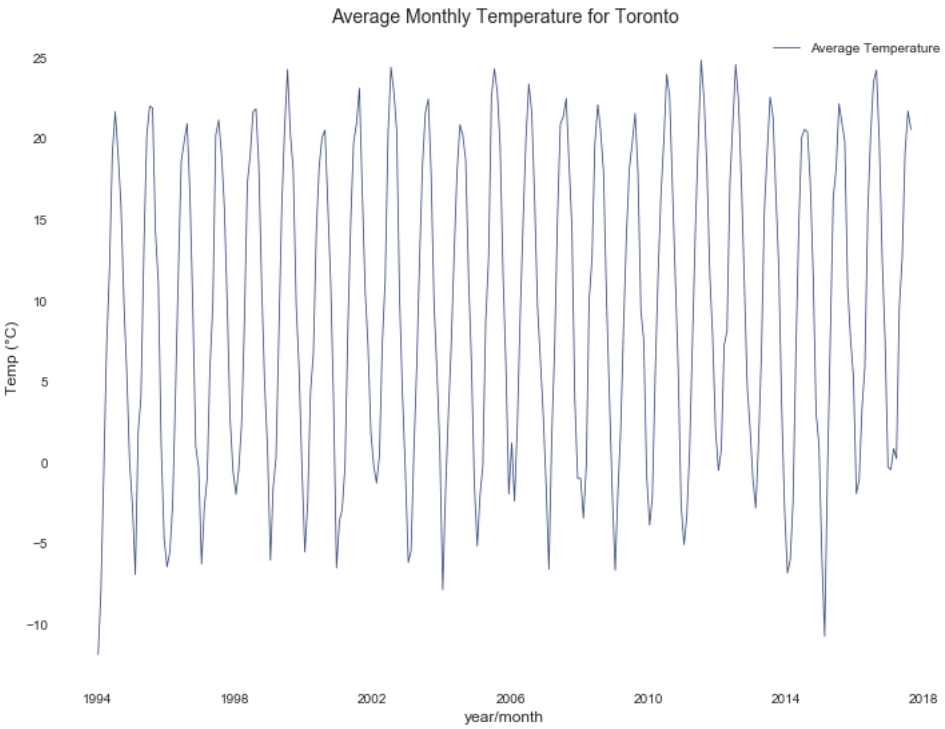
- **Stn Press (kPa)**: is the atmospheric pressure given in kiloPascals.

- **Hmdx**: is an index number used by Canadian meteorologists to describe how hot the weather feels to the average person, by combining the effect of heat and humidity.

- **Wind Chill**: is the perceived decrease in air temperature felt by the body on exposed skin due to the flow of air.

- **Weather**: describes the type of weather. (E.g. rainy, snowy, sunny, cloudy etc.)

We have that the monthly average temperatures have not changed too dramatically over the years. The winters of 2014 and especially 2015 look to be outliers in the recent trend of warmer winters of the past six years. The summers have an average temperature between 20°C and 25°C.



From the IESO we have the following features.

- **Ontario Demand**: The amount of power measured in Megawatts consumed throughout Ontario for that hour.

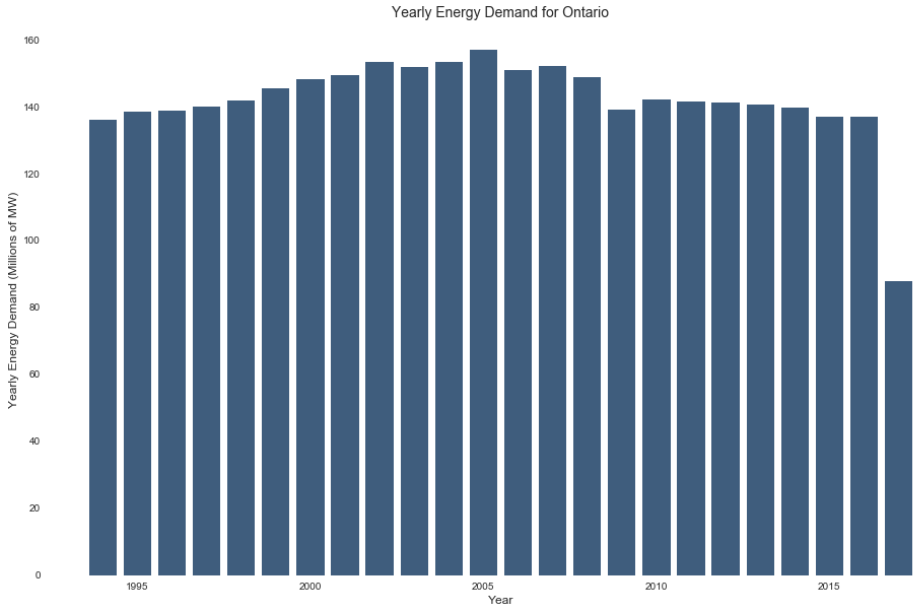
- **Total Market Demand**: is the total energy dispatched into the IESO controlled grid, calculated as Ontario generation plus imports plus generators that have not submitted offers.

- **Imports**: is the total energy injected into the IESO controlled grid from generators outside Ontario.

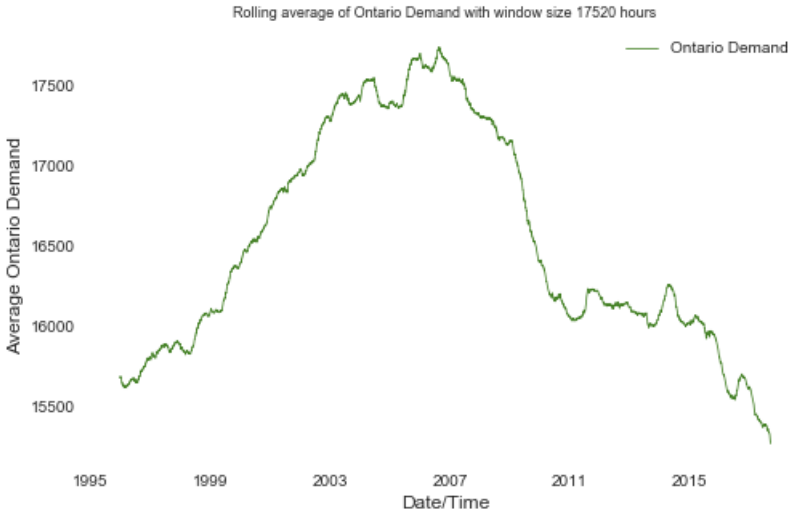
  Exports: is the total energy dispatched outside Ontario from the IESO controlled grid.

  HOEP: is the hourly Ontario wholesale energy market price given in megawatt hours ($/MWh).

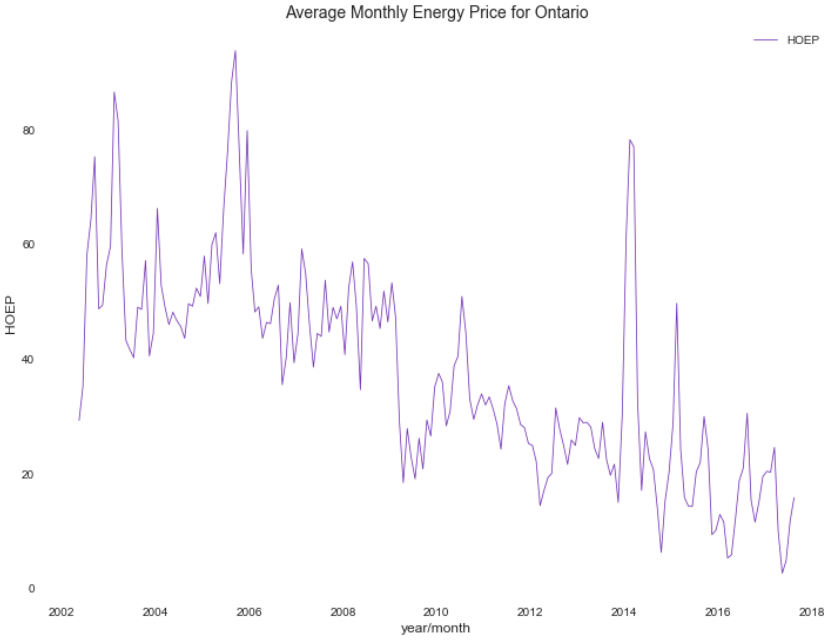
3.2 Energy demand



Before the IESO opened the Ontario electricity market on May 1st 2002, the yearly energy demand for Ontario was steadily increasing. With the exception of 2005, where energy demand peaked, the energy demand of Ontario has been on the decline, even though the population has been steadily increasing. This could be a result of higher prices to consumers or improvements in energy efficiency. The IESO report that coal-fired generation has been phased out; wind and solar generation have joined the provincial supply mix; and new types of demand response and storage resources are also helping to meet the province's demand for electricity. Furthermore, we see that the exports seem to be increasing as well. The drop in demand in 2009 could be a result of the recession. Finally, taking a rolling average with a window size of one year, we see that the average energy demand peaked in 2007 and has been declining ever since.

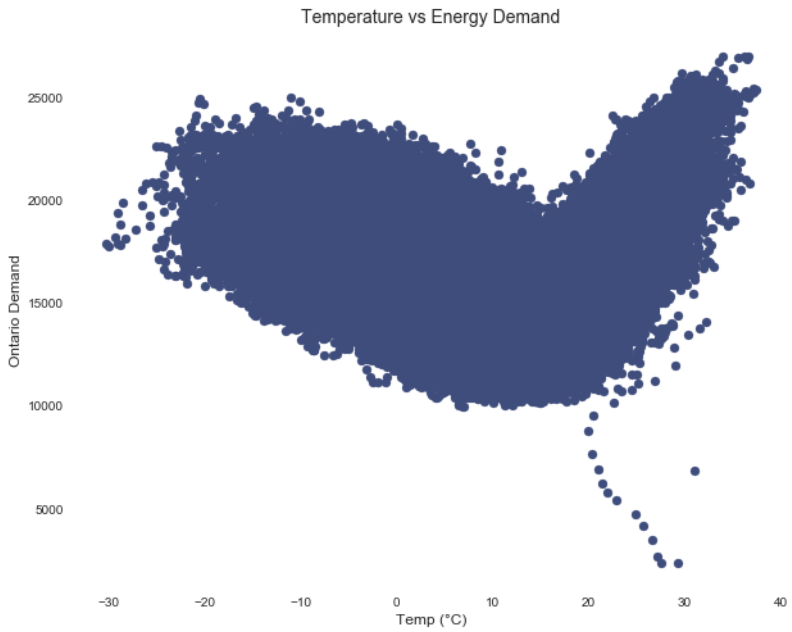


It is interesting to see a general trend of decreasing average energy price, with a few very big spikes in 2005 and 2014, even though average monthly energy demand is falling.



3.3 The relationship between energy demand and temperature

We now look at the relationship between energy demand and temperature.



There is a very clear distinctive relationship between temperature and energy demand. We see that energy demand increases slowly as temperature decreases from 15°C and increases more rapidly as temperature increases from 15 °C. There are also some clear outliers with very low Ontario Demand. On August 14, 2003, Ontario and the north eastern United States experienced a large scale blackout. At the time, it was the world's second most widespread blackout in history. The outliers all took place during a fifteen hour time period on this day.

4 Missing Data

We are missing Ontario energy demand for the whole month of December in 2001 as well as for midnight May 5, 2002, which was when the market first came online. For the month of December to replace the missing data we use linear approximation in intervals of 24 hours to approximate the demand for January of 2002 and for the energy demand for May 5,2002 we use the previous day's demand.

We then use linear interpolation for the Temp (°C), Dew Point Temp (°C),

Rel Hum (%), and Stn Press (kPa) columns when there is only one hour's worth of data missing. When there is more than one hour's worth of missing data we use the previous day's data.There exists a formula for Humidex based on Temp (°C) and Dew Point Temp (°C) that we then use to fill in its missing data.

Finally, there are not many instances for visibility, but 70% of the values are 24.1 km, so we replace the missing values with this value. Similarly, 68% of the values for wind speed are 0 km/h so we replace the missing values with 0 km/h and for the wind direction, we replace the missing values with the mean.

5 Predictive Modeling

Energy demand is a time series analysis problem so we treat it as a supervised regression problem. We try multiple machine learning models like multiple linear regression and random forest regression. The most accurate models we obtain come from a Random Forest Regression model and a Gradient Boosted Machine Tree Regressor.

Model	**Train Accuracy**	**Test Accuracy**
Linear Regression	0.412	0.433
Random Forest Regression	0.626	0.627
Gradient Boosted Regression	0.833	0.83595

5.1 Linear Regression

For the linear regression model we have that pressure is the most important feature with a linear coefficient of -281, which implies that a very small drop in atmospheric pressure gives a significant increase in energy demand. The most important feature which corresponds to a positive relationship with energy demand is Dew Point Temperature with a value of 141.622.

Many of the features in our initial linear regression model had very small values for their magnitude. Thus we performed the KBest feature selection method by computing the ANOVA F-value for the data to reduce the number of features to the 5 most important features. We find that the five most important features are **HOEP**, **Imports**, **Toronto_pop**, **Ontario_pop** and **Exports**. Performing another linear regression with only these features results in an R^2 value of 0.40, but with much smaller coefficients in our original linear model.

5.2 Random Forest

A random forest is an ensemble method that fits a number of decision tree regressors on various sub-samples of the dataset and uses averaging to improve

the predictive accuracy and control for over-fitting. We obtain an accuracy of 62% and investigating feature importance we have:

Rank	**Feature**	**Importance**
1	HOEP	0.711248
2	Temp (°C)	0.171001
3	Imports	0.044065
4	Hmdx	0.041706
5	Toronto_pop	0.013703

Thus, for the random forest model, price (HOEP) is by far the most important feature followed by temperature.

5.3 Gradient Boosting

Gradient Boosting is a sequential technique which works on the principle of ensemble. It combines a set of weak learners and delivers improved prediction accuracy. The learning rate shrinks the contribution of each tree and there is a trade-off between the learning rate and number of estimators. We want to choose a relatively high learning rate. This results in an accuracy of 83% and the most important features are:

Rank	**Feature**	**Importance**
1	Toronto_pop	0.276852
2	Ontario_pop	0.243368
3	HOEP	0.125280
4	Imports	0.067934
5	Exports	0.062903

With the Gradient Boosting model we see a very large increase in accuracy again from the standard random forest regression model. The value of R^2 now has an accuracy of 83.7%. From our cross-validation step it does not seem to be that the model is over-fitting. The other big change is that Ontario population and Toronto population are now the two most important features with price being significantly reduced and temperature also reduced. This is somewhat surprising since there did not seem to be much correlation between the steady increase in population and the slight decline in energy demand over time.

5.4 Bagging

Bagging is used for parallel ensemble models where each model is built independently and the aim is to decrease variance, not bias. We now perform a weighted average from our three main models above: linear regression, random forest and gradient boosting. We assign a higher weight to the gradient boosting model of

0.7 since it was more accurate, a weight of 0.2 to random forest and assign a lower weight of 0.1 to the linear regression model since it was the least accurate. We obtain a weighted average prediction based on the three models of 80.4%. While this model is very accurate, it does not allow us to say anything about the feature importance based on this prediction.

5.5 Conclusion

From our investigation into energy demand, we saw a surprising effect of the introduction of energy production in Ontario to the market in 2002. Demand for energy initially increased and has been declining slightly since 2005, while average monthly exports has been steadily increasing. Along with the expected daily and weekly energy trends, energy demand exhibits clear seasonal trends that were evident in the non-linear relationship between energy and temperature.

The best performing models were the Random Forest Regressor and Gradient Boosted Regression with R^2 values of 63% and 83% respectively. Note, that in the case of the GBM regression model, using feature selection to reduce the number of features resulted in a very small reduction in R^2. The most important features differed quite a lot between the two models with price playing a much more important role in the random forest model whereas population played a very important role in the GBM model, temperature played only a much smaller role than expected, possibly due to its non-linear relationship with energy demand.

In the future it would be nice to incorporate more economic features like whether the economy is in a recession or a boom, housing and oil prices as well as employment data to see how these features are related to energy demand.

6 Recommendations

- Companies monitoring energy consumption should absolutely take weather and population into consideration. Temperature plays a delicate role in the demand for energy with energy demand increasing slowly as temperature decreases from 15°C and increasing more rapidly as temperature increases from 15 °C.
- Both average energy demand and average energy price are trending downwards so Ontario should prepare to export more of its energy in the future.
- For energy demand, it would be interesting to incorporate energy sources into the prediction model. As Ontario increases the percentage of its energy production coming from solar and wind sources, these are very much more tied to weather than previous sources like Nuclear or hydro. Access to this data will hopefully be explored further in the future.