ORIE 4741 Project Final Report

Grace Mitchell(gmm93), Bingyi Fan(bf323), Alex Leong(adl88)

I. Introduction and Objective

In this project, our objective is to predict hourly electricity demand, based on previous electricity demand, current weather factors, and time indexes. This project will attempt to determine which weather factors significantly contribute to electricity demand, and whether the same forecasting techniques can be used to accurately predict electricity demand in different building types.

Specifically, we will try to relate weather in New York City and electricity demands of four buildings of different types in the city. On a large scale, predicting electricity demand is important for electrical distribution grids worldwide to handle the needs of their consumers. Businesses need to be able to budget for their own electrical use. Traditionally, forecast of electricity demand has been based on seasonality and day/night shift. This forecast model is premature and inaccurate in many cases. Understanding how weather factors, as well as additional features, contribute to electricity demand will render a more precise model and facilitate the optimal allocation of resources. Forecasting energy demand can also identify when buildings are operating inconsistently, or using excessive amounts of energy.

The dataset involves a linear relationship between weather factors and electricity demand. For example, as average humidity and temperature increase, we can expect that electricity demand will increase. The accuracy of the machine learning implementation will be relatively simple to evaluate, as it is a supervised learning problem with continuous output. The robustness of the model will also be tested, as it could be implemented across four different datasets, which include two different building types.

II. Datasets

To achieve our goal, we will use data from two time series datasets. The first dataset is weather data collected at Central Park station in NYC, provided by National Center for Environmental Information. Central Park is located near the geographical center of Manhattan Island, and we will assume that the weather at that station is representative of weather of the whole city. The weather dataset contains the following features which we believe are closely related to electricity demand: Dew Point temperature, Dry Bulb temperature, humidity, precipitation and wind speed. The second dataset is provided by Grace Mitchell, one of our team members. It contains time series electricity demands of a courthouse, an office building and two school buildings in the city. These are actually four datasets, as we will train four different models on each type of building. Each time series contains one-year data ranging from 7/1/2016 to 6/30/2017, and we will have all of them to have a uniform time interval of an hour. This is to ensure that time series from two datasets match each other and that they can be safely concatenated. We will build preliminary models on these datasets. We deleted all time information from the dataset, while time in a day still remains as a feature, in an effort to avoid working with time series. Date and time will be considered in aspects of feature engineering, which will be discussed in the next section.

All data from December 8th and 9th was omitted due to a relatively large amount of missing values in weather data. Another problem we encountered when cleaning up the data was that weather data was collected more frequently than one hour and was not “on the hour” for every day. The number of data

points collected was inconsistent for each day, as the weather data was not always collected at the same time, or the same number of times every day. As a workaround, we decided to use data points collected that were nearest to "on the hour" (i.e. 12:51, 1:51, ... 11:51) for each day. We do not believe that this will have a substantial consequence on prediction accuracy, because weather generally does not change in just a few minutes. Because of the inconsistent time interval, a total of 12 data points were missing in the weather data, so 12 additional points were randomly duplicated. For each missing weather data point, the values were interpolated between previous and successive moments in time. Generally, many data points for precipitation and wind speed were missing, but this only occurred in approximately $\frac{1}{4}$ of the data set.

III. Preprocessing and Feature Engineering

The weather features include the Dew Point temperature, Dry Bulb temperature, humidity, precipitation and wind speed for every hour. The electricity demand feature includes the previous hour's electricity demand for every hour. This feature was included in the models, as the objective of this project is to predict electricity demand for specific buildings. This is a way to allow the models to learn the historical electricity demand of that specific building before trying to predict the demand. For example, if you wanted to predict the next 24 hours of electricity demand for the office building, you would input the previous 24 hours of electricity demand into the model trained for the office building. This was inspired by the research paper, "Forecasting Energy Demand in Large Commercial Buildings Using Support Vector Machine Regression.

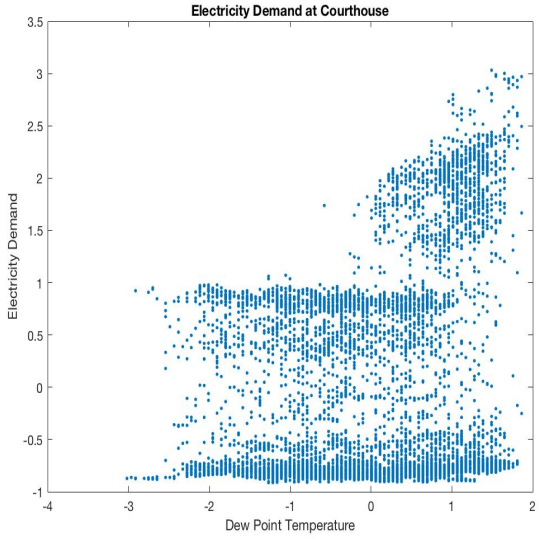

Fig. 1: Electricity Demand by Dew Point Temperature

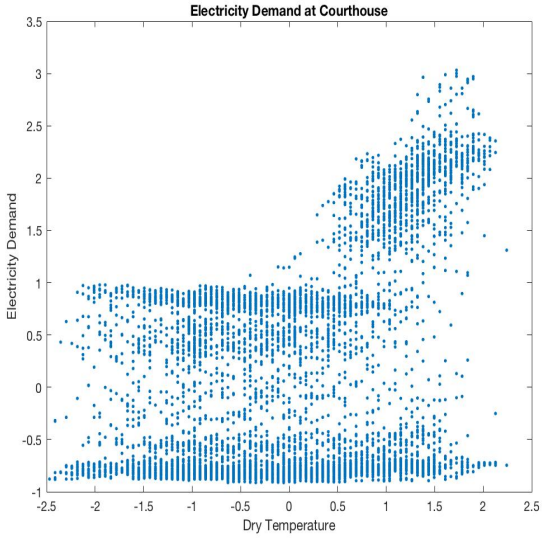

Fig. 2: Electricity Demand by Dry Bulb Temperature

For the time index features, categorical features and a numerical feature were added to the input space. The numerical feature, a time index, included an array of equally spaced values ranging from 1 to 24. These values were aligned with the corresponding hour of the day, for each day within the dataset. The first categorical feature represents the time of day, split into 5 categories. The hours of 8:51PM-3:51AM were considered "Night," the hours of 4:51AM-6:51AM were considered "Early Morning," the hours of 7:51AM-10:51AM were considered "Morning," the hours of 11:51AM-3:51PM were considered "Afternoon," and the hours of 4:51PM-7:51PM were considered "Evening." This categorical time of day feature is based on the typical operation schedule of commercial buildings. Both of these time

index features were included in the input space in the hopes of capturing some behavioral aspects of the building operation patterns that occur at different times of the day.

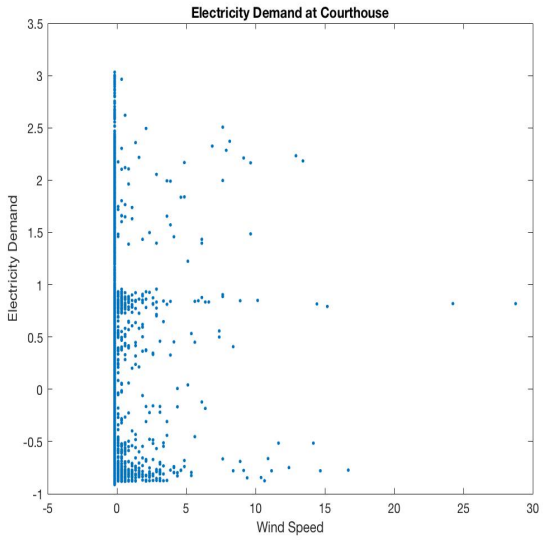

Fig. 3: Electricity Demand by Wind Speed

The second categorical feature accounts for seasonality. The data corresponding to 7/1/16 to 9/15/16 and 5/16/17 to 6/30/17 were considered "Summer." The data corresponding to 9/16/16 to 11/15/16 was considered "Fall," 11/16/16 to 2/28/17 was considered "Winter," and 3/1/17 to 5/15/17 was considered "Spring." The seasonality feature was included in the input space in the hopes of capturing some behavioral aspects of the building operation patterns that occur during different seasons. For example, electricity usage is expected to be higher in the summer due to increased cooling loads.

The third, and final, categorical feature accounts for the day of the week for each data point. The categories are either "Weekday," or "Weekend." The day of the week feature was included in the input space in the hopes of capturing some behavioral aspects of the building operation patterns that occur on weekdays vs the weekend. For commercial buildings, weekend loads have different patterns, as well as significantly lower demand.

All categorical vectors were input in the feature space with the use of one-hot encoding. The total input space included 19 feature columns, one as an offset, two for the day of the week categorical feature, four for the seasonality categorical feature, five for the time of day categorical feature, five for the weather features, one for the hour index numerical feature, and one for the previous hour's demand numerical feature.

All feature vectors were standardized by subtracting column mean and dividing them by column standard deviation. Then, data was randomly divided into three sets, with 60% of it put into the training set, 20% put into the validation set, and the remaining 20% put into the test set.

IV. Models

This project includes the fitting and analysis of three different models: Support Vector Machine Regression (SVMR), Random Forest, and linear regression with k-sparse regularization. Each model, and its results, will be discussed in their own section. The conclusion of the report will include our determination and discussion of the best model out of the three.

We use two indicators to evaluate the results of a model, root mean square error (RMSE) and prediction accuracy based on 10% difference between the ground truth and predicted demand. RMSE is a commonly used value to evaluate a model; however, the energy industry cares more about accuracy, because higher accuracy means more stable predictions. The industry requires accuracy to ensure that power plants are running at the appropriate level of output power, both to meet users' demands and to protect the grid. They do not necessarily care about a small error in prediction.

A. Model 1: SVMR

The use of SVMR was inspired by the research paper, "Forecasting Energy Demand in Large Commercial Buildings Using Support Vector Machine Regression." The use of SVM is more popular for classification problems, but can also be used for

regression. The core idea behind SVM modeling is to learn a function by linearly mapping data into a high dimensional input space, and then to fit either a classification boundary, or function, based on similarities between pairs of data points. The way the data is mapped depends on the kernel function you use. The model complexity depends on the parameters: C, epsilon, and the parameters of the kernel function. The scikit-learn library defines C as the "penalty parameter of the error term," and defines epsilon as the parameter that "specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value."

As there are inherent linear trends in the data (higher temperature corresponds to higher cooling load, for example), the linear kernel function was first used. As the kernel function is defined by the user and different kernel functions generate different results, the polynomial kernel function, of degree 4, was also used. Increasing the polynomial degree from 3 to 4 lead to more accurate results, but it was not increased past 4 in an effort to avoid overfitting. The results of the linear kernel model were compared to the polynomial kernel model, and the one with the best results, or highest average accuracy on the validation and test sets, was selected moving forward.

In an effort to determine the optimal value, and the impact of, the user-defined model parameters, the model fit was iterated for different values of C. Different ranges of C were considered for each model, as different kernel parameters require a variable range of C. The choice of changing C over epsilon was arbitrary. The same iterative method could potentially be used to determine the optimal value of epsilon for either kernel-defined model.

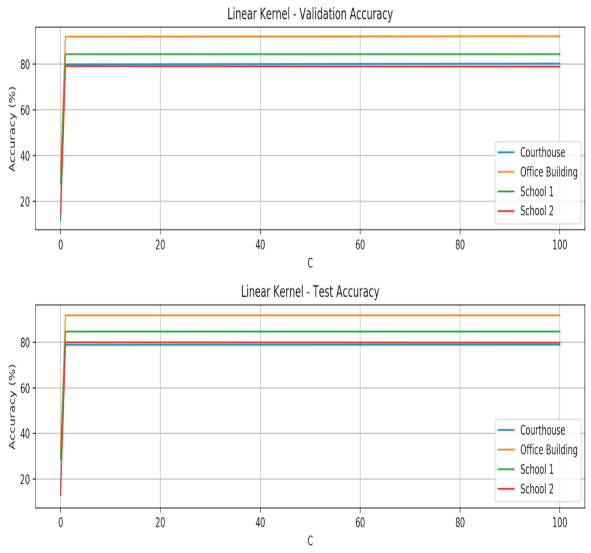

Fig. 4: The plots compare the validation set and test set accuracy, for different values of C, for the model defined by the linear kernel function

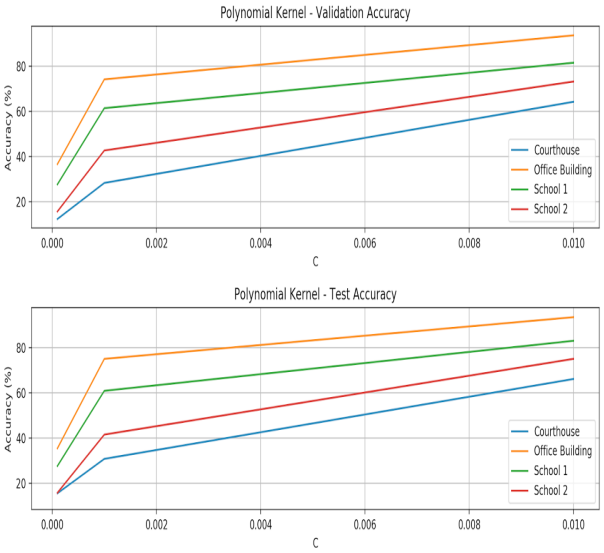

Fig. 5: The plots compare the validation set and test set accuracy, for different values of C, for the model defined by the polynomial kernel function

B. Model 1: Results

The model defined by the linear kernel function was selected moving forward, due to higher accuracy on both the validation and test set. For the linear kernel model, the average validation accuracy was 83.9% and the average test accuracy was 83.8%. For the polynomial kernel model, the average validation accuracy was 78.1% and the average test accuracy was 79.4%.

Linear Kernel SVMR - Results				
	Validation RMSE (kWh)	Test RMSE (kWh)	Validation Accuracy	Test Accuracy
Courthouse	93.245	86.740	79.908%	78.805%
Office Building	26.946	28.977	92.021%	91.729%
School 1	26.365	26.264	84.386%	84.607%
School 2	42.387	41.182	79.105%	79.897%
Average	47.236	45.791	83.855%	83.759%

Fig. 6: The table displays the results of the model defined by the linear kernel function

C. Model 2: Random Forest Regression

The use of random forest regression was motivated by the fact that it can operate on both continuous and categorical features directly. The model generates a number of deep decision trees on the data and outputs the mean prediction of the forest for each prediction. Deep decision trees inherently overfit to their training data, so the use of a forest and the average predictions mitigates this issue. One of the key features of random forests is that for each decision tree at each candidate split in training, only a random selection of features are selected. For data with d features, d/3 features are selected. This random feature selection decorrelates features which are strong predictors, and makes prediction more accurate (Ho, 2002).

The package implements k-fold cross validation, so the training and validation sets were combined during the training stage of the random forest with a value k = 10. The tree count for these random forests started at 10 and was incremented by 10 up to 250 trees. The validation and test RMSE values as well as percentage accuracy are plotted below with respect to each forest size.

It was found by inspection that at around a size of 200 trees, the random forest model is optimized.

The package used to perform random forest regressions was developed chiefly by Henrik Bostrm, computer science professor at the University of Stockholm.

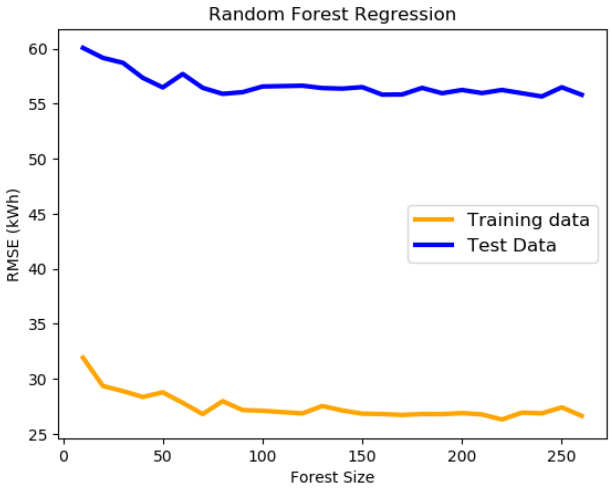

Fig. 7: RMSE Values by Forest Size

Fig. 8: Percentage Accuracy by Forest Size

D. Model 2: Results

These results are from random forest models generated with 200 trees at maximum depth. The models perform well on their test sets, however they exhibit overfitting on their training data.

E. Model 3: Linear Regression and Huber Regression

Two kinds of loss functions, quadratic loss (or, least square) and Huber loss were tried for regression, both with k-sparse regularizer. Since our dataset has 19 features, containing more about 10 kinds of time and weather information, it makes sense to enforce sparsity to reach a simple and practical model for electricity demands forecast in the real

Random Forest Regression - Results				
	Validation RMSE (kWh)	Test RMSE (kWh)	Validation Accuracy	Test Accuracy
Courthouse	26.603	55.924	93.671%	81.619%
Office Building	9.022	20.876	99.870%	96.611%
School 1	8.768	20.249	95.451%	82.940%
School 2	13.122	29.352	93.499%	80.183%
Average	14.378	31.600	95.622%	85.338%

Fig. 9: The table displays the results of the random forest regression model

world. Quadratic loss is the most commonly used loss for linear regression, so it serves as a baseline model. Huber regression tends to create a robust regression. The results will show that quadratic loss usually has smaller RMSE and lower accuracy, while Huber has slightly bigger RMSE, but much better accuracy. As mentioned before, the energy industry wants stable predictions, so we think that Huber regression may be preferable here. Two models, with different k values for sparsity, are compared with each other and a best model is chosen for each type of building. A relatively good k was chosen for each model based on accuracy and RMSE, and reported below.

	Mean (kW)	Standard Deviation (kW)	Median (kW)	Max (kW)	Min (kW)
Courthouse	496.3985	325.6817	297	1470	200
Office Building	463.1349	161.3053	425	1029	270
School 1	176.1915	84.6935	164	525	58
School 2	240.6692	147.3885	175	772	77

Fig. 10: Statistics of the “ground truth” electricity demand of the four buildings

F. Model 3: Results

8-sparse works well for both regressions. The most important features are previous electricity demand, day type and time in a day. When forcing sparsity, these features almost always show up, while weather features only contribute a little to refining the model.

Least square find the mean and Huber find the median. Electricity demands often have irregular distributions, so in general, Huber regression is preferred than least square when predicting electricity demands.

Least Square Regression (Quadratic)					
	k-sparse	Validation RMSE (kWh)	Test RMSE (kWh)	Validation Accuracy	Test Accuracy
Courthouse	8	89.1185	82.8968	59.30%	58.47%
Office Building	8	27.8209	29.6106	89.95%	90.92%
School 1	8	25.7011	25.3626	75.66%	75.36%
School 2	8	40.1024	39.6931	56.43%	56.98%
Average	-	45.6857	44.3908	70.34%	70.43%

Fig. 11: The table displays the results of the least squares regression model

Huber Regression					
	k-sparse	Validation RMSE (kWh)	Test RMSE (kWh)	Validation Accuracy	Test Accuracy
Courthouse	8	98.5791	92.6127	82.49%	82.02%
Office Building	8	30.6659	33.7845	91.91%	89.83%
School 1	8	26.5661	26.3597	83.98%	84.55%
School 2	8	43.4854	42.1099	79.05%	79.78%
Average	-	49.8241	48.7167	84.36%	84.04%

Fig. 12: The table displays the results of the huber regression model

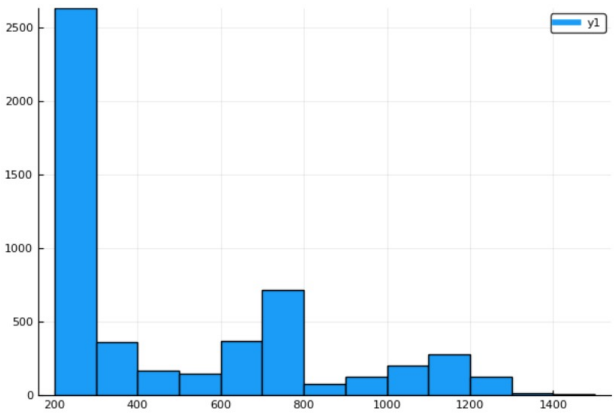

Fig. 13: Y of courthouse

The courthouse has a very skewed distribution so Huber regression works much better than least square.

V. Conclusion

When comparing the values of average accuracy for the validation and test sets, the random forest is the best, most accurate predictive model. The Huber regression and the SVMR models had less accurate, but similar results. When comparing the values of RMSE for the validation and test sets as well, the random forest remains as the most accurate predictive model.

As random forest models typically have issues with overfitting and each building type had varying performance with each model, we recommend that the company use the best model fit for the building type when considering predictive analysis. The difference between the validation accuracy and the test accuracy, or the generalization gap, is wider for the random forest model. For both the Huber regression and SVMR model, the generalization gap is much smaller. To have sufficient confidence in predictions of electricity demand, it is recommended to use both the random forest and either the Huber regression, or the SVMR, model and average the predictions. For a courthouse, it is recommended to use both the random forest model and the SVMR, as the SVMR has the second lowest RMSE across both sets. For an office building, it is recommended to use both the random forest model and the SVMR, as the SVMR has the second highest accuracy and second lowest RMSE across both sets. For a school, it is recommended to use both the random forest model and the SVMR, as the SVMR has the second highest accuracy and second lowest RMSE across both sets.

Based on the model results, we could predict electricity demand with approximately 76-80% confidence. We would be willing to use them in production to change how our company makes decisions.

VI. Reference

- David M. Solomon, Rebecca Lynn Winter, Albert G. Boulanger, Roger N. Anderson, Leon Li Wu, 2011, Forecasting Energy Demand in Large Commercial Buildings Using Support Vector Machine Regression, Columbia University Academic Commons, https://doi.org/10.7916/D85D90X7
- Henrik Bostrm, P 2017. Julia implementation of random forests for classification and regression with conformal prediction, from https://github.com/henrikbostrom/RandomForest
- Ho, Tin Kam (2002). "A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors" (PDF). Pattern Analysis and Applications: 102?112.
- Scikit-learn developers. (2007). Sklearn.svm.SVR. Retrieved December 4, 2017, from http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html