FINAL REPORT - TEAM POWER RANGERS - Energy Usage for Large Buildings within the District of Columbia (DC), 12 Dec 20

Authors - Dominique Hinton, Roman Villoria, and Carla Coulson

**Project Abstract**

The energy sector is a key contributor to climate change, accounting for more than two-thirds of global greenhouse gas (GHG) emissions. And, the energy sector is directly impacted by climate change through increases in usage/consumption of commodities such as water and electricity.

Buildings are significant energy consumers. In 2019, end-use energy consumption by residential and commercial sectors was equal to 28% of total energy consumption in the United States. Our proposed project will predict monthly energy usage based on past energy usage and weather data for a variety of building types, all greater than 50,000 SF in size, located within DC. Given current climate change concerns, having an accurate estimate of future energy usage in large buildings will help the District, and utility providers, effectively conduct strategic planning required to optimize resources, better estimate building operating costs, and set energy and GHG emissions reduction standards and renewable energy goals.

**Initial Hypothesis**

Our primary hypothesis: Given energy usage data for buildings over 50,000 Gross Square Feet in the District of Columbia, as well as facility Energy Star (ES) ratings and weather data, we can predict future energy usage in a manner that may generate savings for building owners (or operators).

Other initial thoughts and considerations included:

* The best performing buildings are those most recently constructed and with higher ES ratings.
* Buildings with improving ES ratings over time exhibit reduced energy and water consumption and GHG emissions.
* Water usage is correlated to overall energy usage.
* Increases in energy and water usage are associated with extreme weather evidenced by climate change.
* Changes in energy usage due to the COVID pandemic have resulted in decreases in energy and water usage.

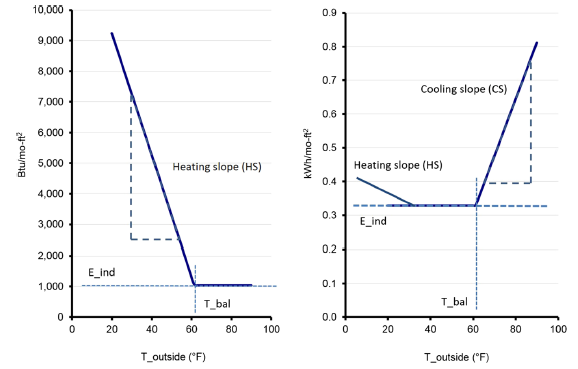
As our dataset included water usage as an annual versus monthly value, our focus throughout the project was building energy usage.

**Amended Hypothesis**

As we progressed through our Data Science Pipeline, our initial hypothesis was amended as follows: Can we predict DC large commercial buildings’ energy usage based on their historical electricity and natural gas consumption and weather patterns?

The figure below provides expected slopes for BTU/month-ft2 of energy usage for cooling and heating seasons, using natural gas (left hand side) and electricity (right hand side) respectively. Energy usage in DC buildings with areas greater than 50,000 SF should follow these patterns.[[1]](#footnote-0)

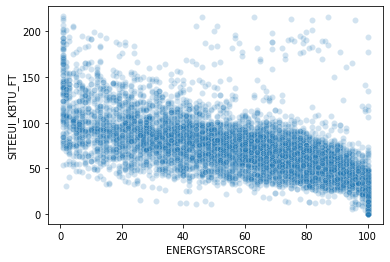
FIGURE 1 - Energy Usage Slopes - Heating and Cooling Seasons



**Applications**

The predictability of energy usage, given weather data, ward and building type will provide the District, and utility providers, with information they can use to move forward with effective strategic planning to optimize resources, better estimate building operating costs, set energy and GHG emissions reduction and renewable energy goals. Also, this information could assist building owners in making decisions on capital improvements that increase energy efficiency (as measured by ES ratings). We observed in our Exploratory Data Analysis (EDA) that higher ES scores are somewhat correlated with lower energy usage as shown in the scatter plot below.

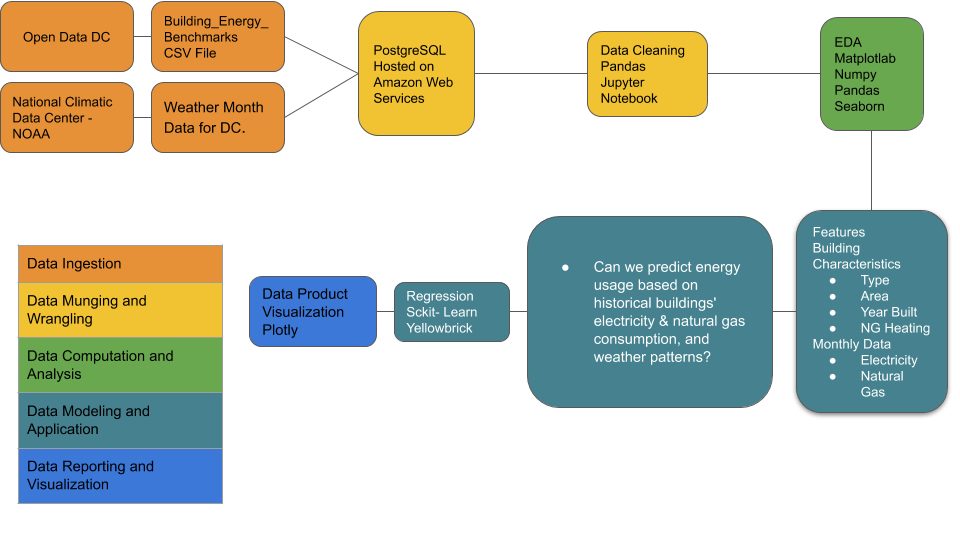
FIGURE 2 - Energy Star Scores for varying SITE EUI (KBTUFT)



More information on ES scores can be found here.[[2]](#footnote-1)

**Initial Data Architecture and Ingestion**

FIGURE 3 - Data Architecture



Our initial data architecture (above) varies little from our final approach, which uses various regression models to predict energy usage.

We used two data sources for our Capstone project as follows:

1) Energy Benchmark Data from the DC Department of Energy and Environment (DOEE). This dataset has 71 columns and provides a variety of building characteristics and energy use data for large buildings from 2010 through 2019. Data is updated routinely; a link to the dataset provides easy access to current data. We established an interface (URL to csv file) to ensure availability of the latest DC buildings’ usage data.[[3]](#footnote-2) Our Capstone project focused on data from 2018 and 2019 as it was the most complete.

In addition to this dataset, we tried to add more information from other datasets. However merging datasets by building-month was not possible. Other datasets explored included:

a) Historic\_Data\_on\_DC\_Buildings.csv: <https://opendata.dc.gov/datasets/8ffa9109cd9a4e37982cea67b289784d_0?geometry=-77.738%2C38.814%2C-76.286%2C39.001>

b) Monthly renewable energy data for DC commercial buildings: data not available for DC

c) https://opendata.dc.gov/datasets/green-buildings-leed: Green\_Buildings\_\_LEED\_.csv: Only 8 buildings matching in DC

d) https://emp.lbl.gov/tracking-the-sun: Data for Washington has too many null values

e) https://www.taxpayerservicecenter.com/RP\_Search.jsp?search\_type=Assessment: Address, tax Class (example: Commercial), Total Value

f) https://www.energy.gov/eere/buildings/building-performance-database-bpd

g) https://bpd.lbl.gov/explore

h)https://energy.duke.edu/research/energy-data/resources?f%5B0%5D=field\_dataset\_topic%3A279

i) https://www.nrel.gov/buildings/comstock.html

j) https://opendata.dc.gov/datasets/computer-assisted-mass-appraisal-commercial

2) DC weather data and associated characteristics were initially acquired from OpenWeather. The team also downloaded historical data from National Oceanic and Atmospheric Administration (NOAA) and compared the two datasets to determine which was the most simple and elegant to use. Since the weather data available at OpenWeather required significant transformation, from hourly to monthly data points, the team decided that the NOAA dataset, which provides an array of weather characteristics as monthly information, best fit the needs of the project.[[4]](#footnote-3)

Power Rangers established a PostgreSQL database in the Amazon Web Services Relational Database System. This option was chosen so we could implement the WORM storage method into our pipeline. At first, the team was moving in the direction of utilizing a SQLite database. This was determined to not be the best course of action because of the potential for accessibility problems. We then explored using other open source and server-based SQL clients. The question of how to securely access the database was explored, as the team did not want the credentials to be hard coded into a Jupyter notebook. As such, we deemed that use of a configuration file was feasible and appropriate.

The PostgreSQL database contains data as described earlier. The database tables/schema were set up within the database connection practice notebook located within the Power Rangers repository. The raw data was inserted using the pgAdmin GUI. After data cleaning and use of several different methods to load the datasets, this was determined to be the optimal way to move raw data to the database. The team is using the SQLAlchemy create\_engine function to connect to the database and pull data into a Jupyter notebook, and the data frame function to\_sql to create tables in the database. The establishment of this database was extremely time consuming, however, it gave the team the ability to work elements of the data science pipeline in SQL.

**Data Wrangling and EDA**

The python library of tools (Pandas, Numpy, Matpoltlib, Seaborn, Scipy) was used to clean and wrangle the data and to conduct EDA.

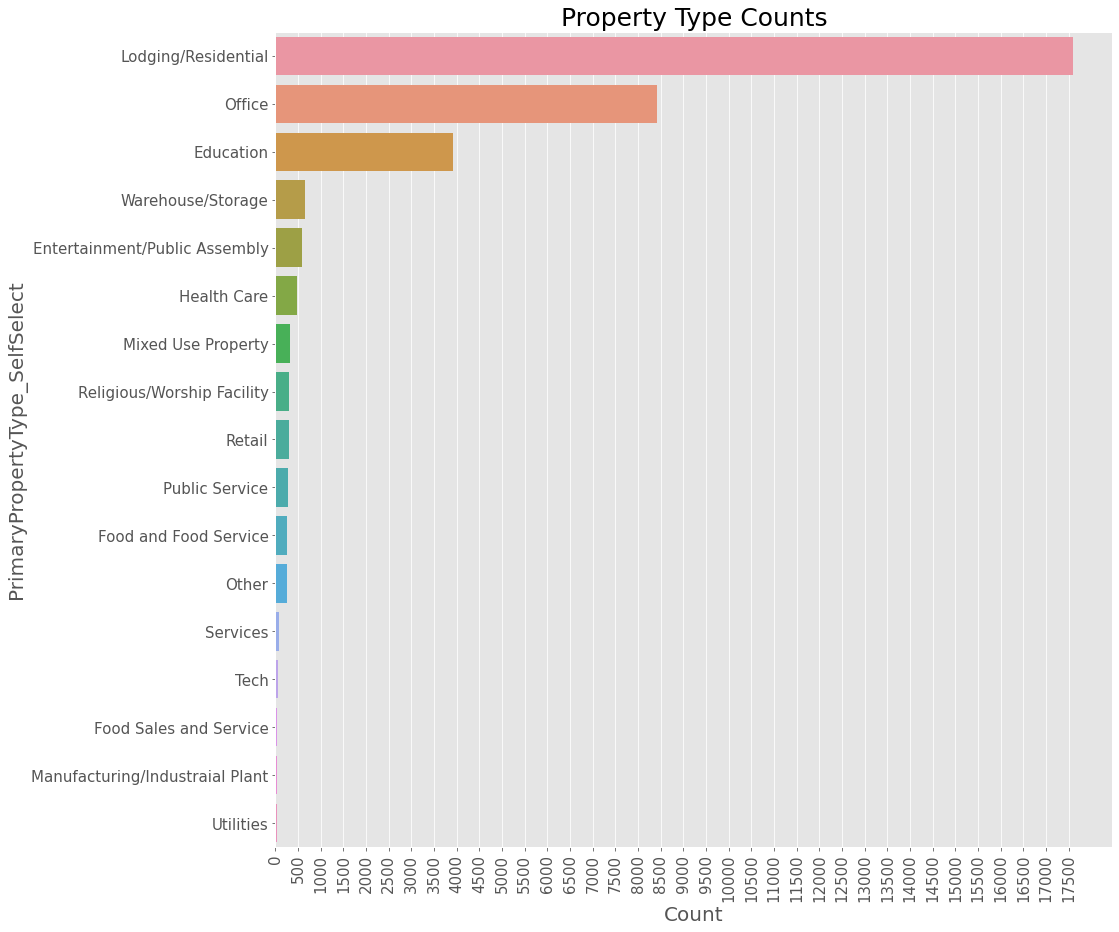
Power Rangers inspected monthly energy usage records to: obtain counts of buildings with 12 months of valid usage for each year, identify duplicated building instances (if any) and to identify building-year combinations with null usage.​ We also inspected and counted annual data such as water usage, renewable energy usage, GHG emissions and ES scores across variables grouped by reporting year and property type. The team plotted histograms of buildings by reporting year and ES score and produced descriptive statistics for ES scoring changes over time. We explored trends between ES scores and energy usage (site Energy Use Intensity(EUI)), ES scores and Renewable Energy usage, and energy usage and year built, with scatter plots. We looked at site EUI by year and building type and produced a correlation matrix between several key variables. ​

As part of our initial EDA, we met with representatives from the DC DOEE Energy Benchmarking Team. The meeting was invaluable in that it provided us with a better understanding of the data and gave us additional information essential to our project. For example, it was explained that the District works hard to ensure their energy benchmarking data is valid, complete and accurate. They shared that the algorithm used for Energy Star scoring was revised between 2017 and 2018 and we should observe a decrease in scores when comparisons are made between those years. We also learned that 2020 energy benchmarking data will not be available until the spring of 2021, and the 2019 data continues to be updated.[[5]](#footnote-4)

Our EDA helped us better understand the energy benchmarking data set. We identified null values; data completeness is particularly important for electricity and natural gas usage information. We observed that the later the year, the more robust the data. We confirmed this with the DOEE. Data was cleaned; rows with null values eliminated. Team Power Rangers concluded that the data from 2018 and 2019 was best suited for analyses and used that splice of data for further work in this Capstone project.

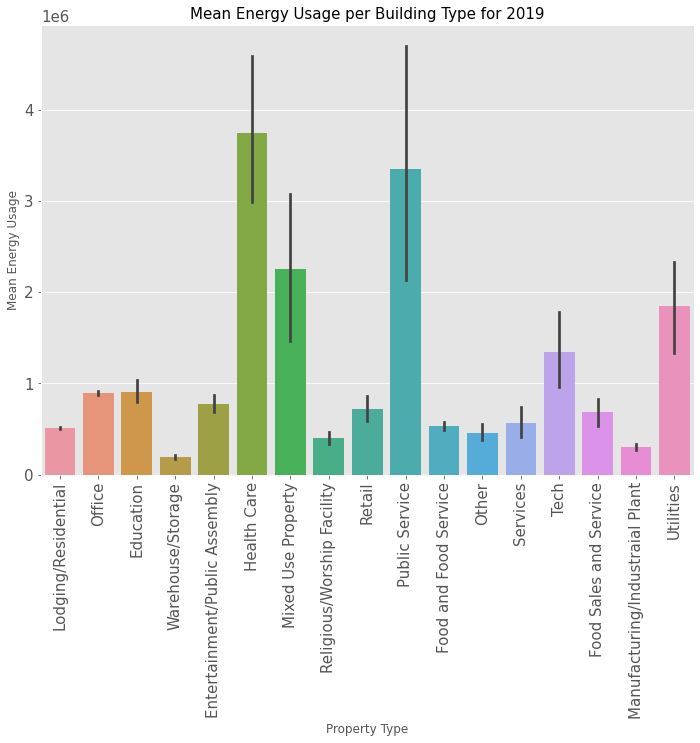
An essential feature of our data is property type. Upon analysis, we determined it essential to group property types to decrease the number from 70 to 19 types. We used property type groupings provided by the DOEE to crosswalk the larger number of types to a more manageable set.​ The groupings that are used throughout our project follow below.

FIGURE 4 - Instances (building month) by Building Type (as modeled)



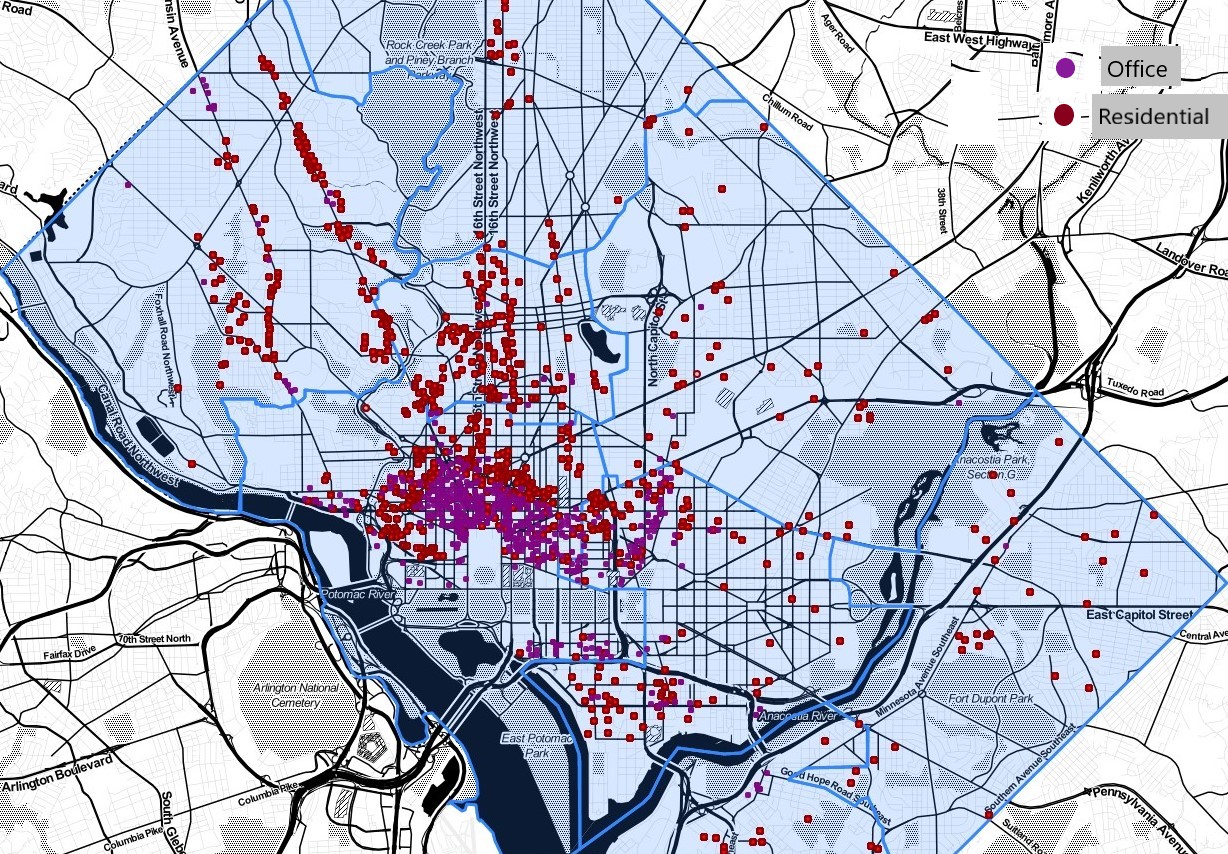
Small instances of certain building types could not be modeled individually, e.g. Banking /Financial Service (0 instances), Manufacturing/Industrial and Parking. Lodging/residential and office building types make up about 80% of our instances. In addition, looking at the mean energy usage/building type for 2019, we see that the smallest variability in energy usage for these two building types.

FIGURE 5 - Mean Energy Usage per Building Type for 2019



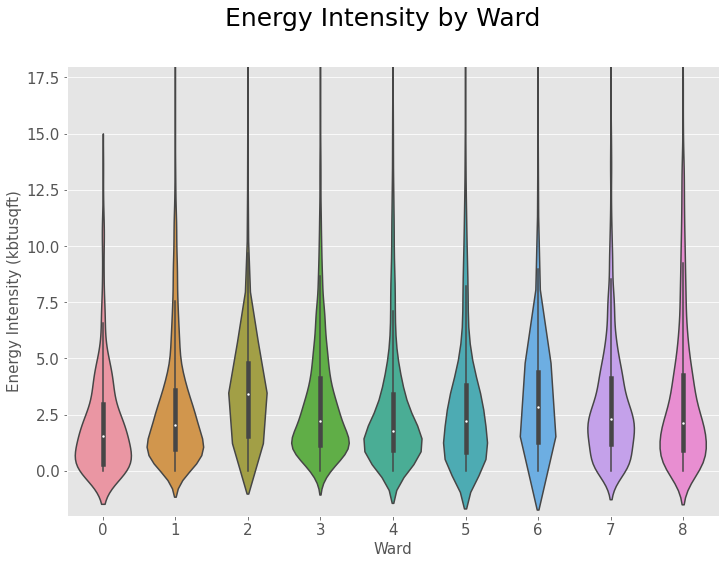
In reviewing the location of our buildings by ward, we observed the highest density of lodging/residential and office building types in Ward 2, which comprises the area that includes Chinatown, Downtown, Dupont Circle, Foggy Bottom, Georgetown, the National Mall, Penn Quarter, the West End and Logan Circle.[[6]](#footnote-5) See Figure 6 below.[[7]](#footnote-6) This area is a rich target to model, as it has many of the building types with the largest number of instances and the smallest variability of energy usage.

FIGURE 6 - Lodging/Residential and Office Building Types by Ward



In this dense area of our city, it is likely that temperature plays a significant role in increased observations of high EUI as compared to the rest of the city. We did see that Ward 2 has the highest energy intensity of the 8 DC wards, as shown in our violin plot below. The DC DOEE should consider focusing their energy use reductions efforts there where we are seeing highest building density and highest energy intensity.

FIGURE 7 - Energy Intensity by DC Ward

Our EDA also provides that:

1) The highest EUI appears to be at Data Centers (of which there are few) and utilities facilities such as Drinking Water Distribution facilities and Wastewater Treatment Plants, also few instances.

2) Renewable energy usage is so small that it is unlikely to significantly affect outcomes. We later dropped this as a feature. We also dropped annual measurements for water usage due to a lack of annualized data.

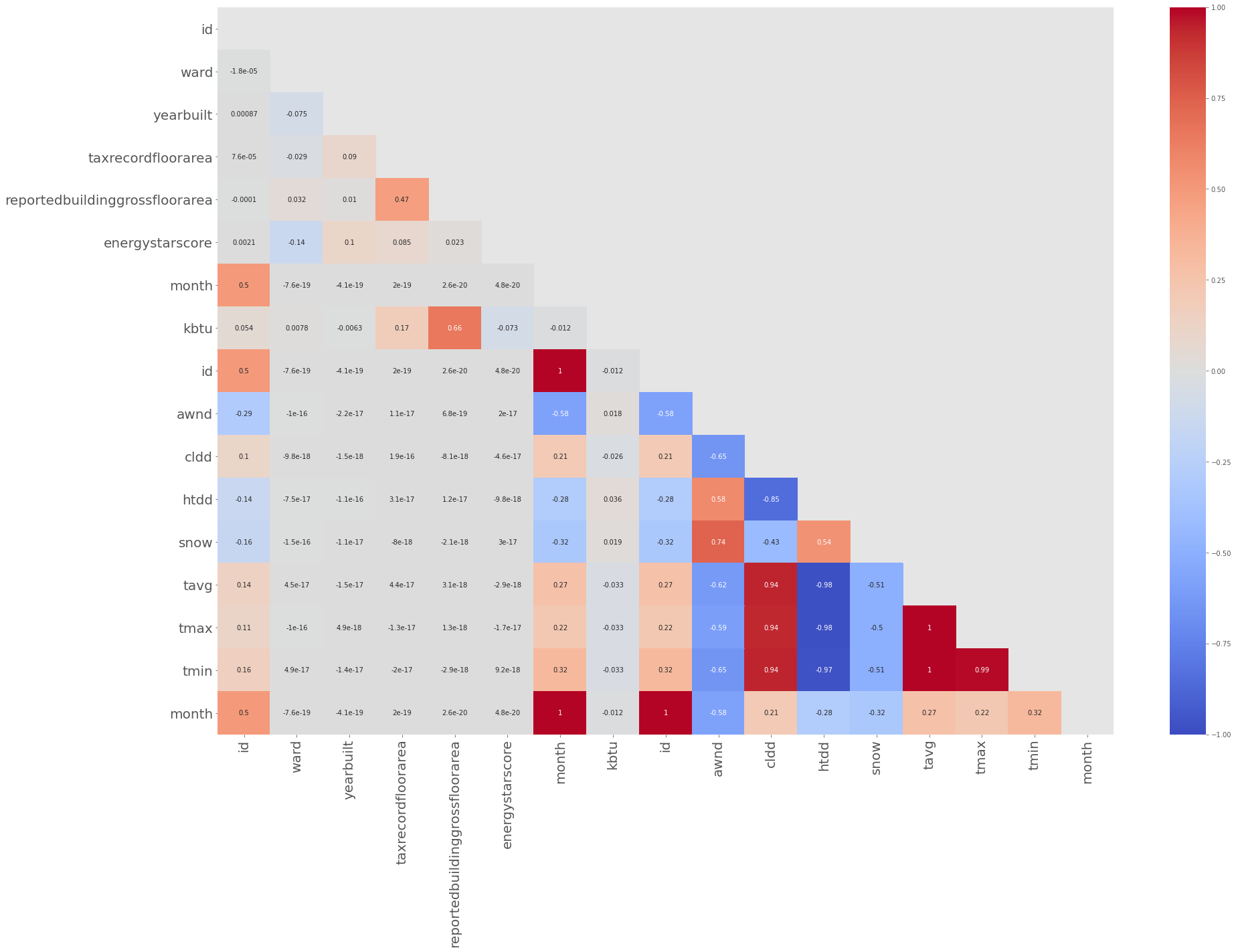
3) Energy Star count increases in later years. Scores are reset in 2018 as evidenced by score decreases between 2017 and 2018.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **FIGURE 8 - ENERGY STAR SCORE (2013 - 2019)** | | | | | | | |  |
| **Reporting Year** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **2013** | 738 | 56.8 | 29.15 | 1.0 | 33.0 | 63.0 | 81.0 | 100.0 |
| **2014** | 1070 | 60.4 | 28.7 | 1.0 | 39.0 | 67.0 | 83.0 | 100.0 |
| **2015** | 1162 | 62.8 | 28.1 | 1.0 | 42.0 | 71.0 | 85.0 | 100.0 |
| **2016** | 1184 | 63.9 | 27.2 | 1.0 | 44.0 | 71.0 | 86.0 | 100.0 |
| **2017** | 1235 | 64.0 | 27.3 | 1.0 | 43.0 | 72.0 | 86.0 | 100.0 |
| **2018** | 1664 | 59.0 | 27.3 | 1.0 | 40.0 | 64.0 | 80.0 | 100.0 |
| **2019** | 1521 | 60.0 | 26.1 | 1.0 | 42.0 | 66.0 | 81.0 | 100.0 |

4) There appears to be some negative correlation between energy star scores and energy use intensity (as shown in scatter plot on page 1). However, because of ES score reset in 2018, observation of little improvement in scores between 2018 and 2019, many buildings without scores in those two years and little improvement in model scores when included as a feature, ES score was eventually dropped from most of our calculations and analyses.

5) Before we began our machine learning modeling, we built several heatmaps to investigate where our features might be correlated. Below is an example of one of our later efforts. One observation is that energy usage in kbtu is correlated somewhat to reported building area, but not so much for tax reported building area. This, in part, assisted us in determining which area to use in our models.

FIGURE 9 - EDA correlation heatmap for most features.



In conclusion, our data wrangling and EDA efforts provided us with a reliable dataset that we further investigated and transformed during our modeling efforts. Our efforts facilitated a good understanding of the instances and locations of various building types, which use the most energy and the impacts of weather on building energy usage. These steps were an essential part of preparing for machine learning; our confidence in our data preparation and transformation carries over to the confidence we have in our overall results and ability to predict energy usage in DC.

**Model Selection Triple**

Team Power Rangers followed the machine learning workflow of Model Selection Triple[[8]](#footnote-7) to understand and engineer features, build and compare the performance of various algorithms (models) and then tune the best performing to steer model selection and optimize results.

**Initial Feature Engineering**

The team’s feature engineering goal was to reduce model complexity and use only those features contributing to and describing relationships inherent in the data.

We began with 2 datasets, DC Building Energy Benchmarks,[[9]](#footnote-8) which consists of about 70 attributes for buildings over 50,000 Gross Square Feet, and the NOAA weather data, which provided an additional 62 weather attributes for weather at Washington National Airport, measured monthly across an array of years. Our first effort was to look closely at the data and identify if there were missing values, outliers, incorrect values, contrived data or data features that were so closely correlated that they would undermine the statistical significance of an independent variable through multi- or collinearity. We did this throughout our EDA and continued to modify our features during the algorithm selection and machine learning pipeline processes.

We quickly identified that there were only 2 years of building data complete enough to use to model and predict monthly future energy usage - 2018 and 2019. We confirmed this with the DC DOEE in early November and quickly limited our focus to times series analysis across these 24 months. As we determined our building instance to be building month, we concluded with

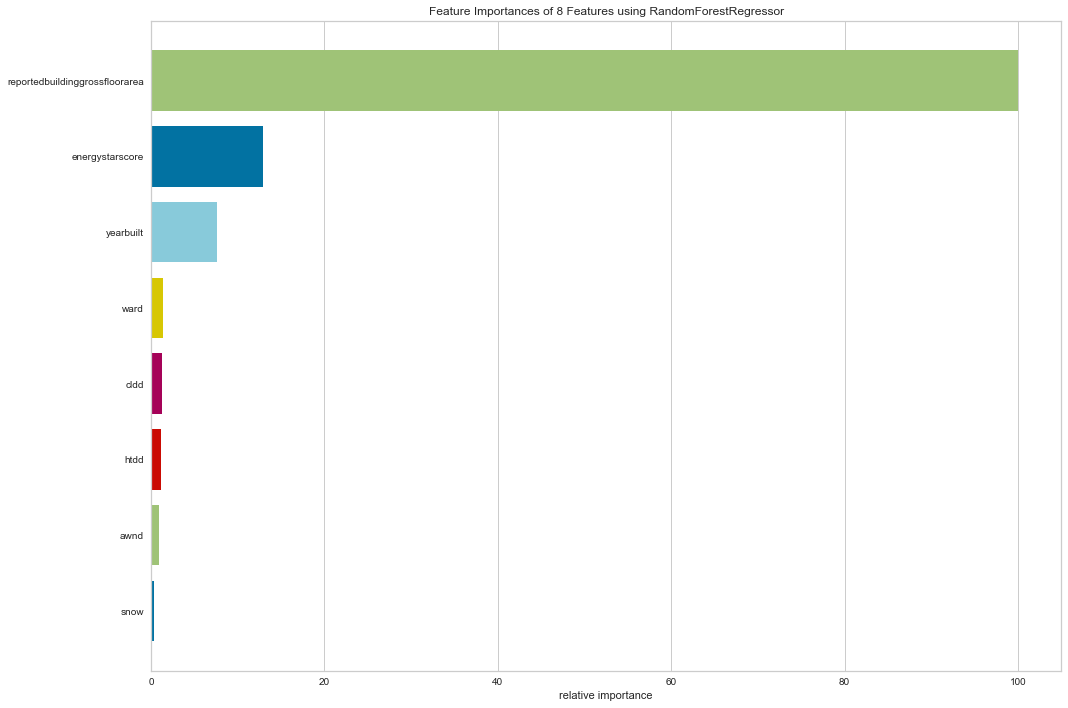
41,304 instances across all building types. We also made the decision to look only at site energy (energy consumed within the building), and discount source energy, which includes the energy required to acquire, refine and distribute the commodity to the site (building). In addition, since our hypothesis focuses on monthly energy usage, we found annual measurements for water usage and GHG emissions to be of little value since they were repeated across instances. We confirmed with the DC DOEE the sparsity of renewable energy data, as well as the reset of ES scores in 2018. And, as described earlier, we also crosswalked building types, given DC DOEE building characterizations, to reduce the number of types from 70 to 19. Finally, we eliminated all weather normalized features as we were using NOAA weather data within all of our modeling efforts and there would have been a degree of multi- and collinearity between these features. Through our initial data investigations we were able to eliminate many attributes that we deemed inconsequential to our problem solving efforts and conclude manual feature construction with a much pared-down list of attributes. This helped reduce complexity as we entered the algorithm selection phase of our Model Selection Triple.

We also devised the attribute of date-time and used it to create primary keys to link together our two PostgreSQL tables, connecting monthly data with reporting and weather years. This attribute was essential to our instance of building month.

Feature engineering became more critical as we began our modeling efforts. As we observed very low R2 scores for all of our models, we changed one feature - “ taxrecordfloorarea” to “reportedbuildinggrossfloorarea”- and saw a vast improvement in scores. Going back and looking at minutes taken at our meeting with the DC DOEE, we noted that the professionals managing the DC Energy Benchmarking program had warned us that building areas reported from tax records were much less reliable than what was self-reported by those filing directly with the District.

In an effort to determine if ES scores can predict improvements in energy usage (ie. less usage), we ran a few of our models using ES Score as a feature. When included in the building type “office” modeling, a Feature Importance scale using RandomForestRegressor shows that the feature is of second most importance behind building area.

FIGURE 10 - Feature Importances of Features using RandomForestRegressor



Of note, the building type, “office”, has the second largest number of instances at 12,864 of 41,304, or about 31% of all instances, the model - Voting Regressor ensemble scored slightly better with ES Score as a feature as shown below. This may not be the case when we eliminate ES Scores that equal zero, or when we apply globally.

Building type ‘Office’, without ES Score (Voting Regressor), R2 mean = 0.88 ; Standard Deviation = 0.04

Building type ‘Office’, with ES Score (Voting Regressor), R2 mean = 0.89; Standard Deviation = 0.02

We entered the algorithm selection phase of the Model Selection Triple using 17 features from the initial 132 available to us from our datasets. Our feature engineering efforts helped us reduce complexity and likely contributed vastly to the overall quality of our results.

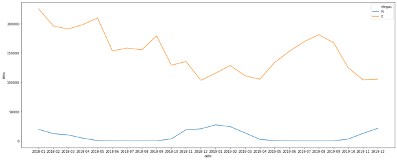
**Algorithm Selection (and continued Feature Engineering)**

The hypothesis for our modeling efforts was - Can we predict DC large commercial buildings’ energy usage based on their historical electricity and natural gas consumption and weather patterns?

Improving building sector energy efficiency is an essential target to reduce GHG emissions, as well as fossil fuel consumption. One of the most effective approaches to reducing emissions and energy consumption, with respect to new construction, is to consider energy efficiency measures during building design. And, efficient energy management and smart refurbishments may also enhance energy performance of existing facilities. These solutions, collectively, require accurate prediction of energy usage to optimize building owner decision-making.[[10]](#footnote-9)

To test our hypothesis we used publicly available data to model energy usage across a variety of building types and its relationship with weather conditions. Initially, we decided to model electricity and natural gas consumption separately to account for the fact that different fuel types cover uniquely different building loads. For example, electricity is used for baseload, heating and cooling, while natural gas is used for baseload and heating, but not cooling. In this sense, we plotted a time series chart to check for trends and seasonality. The figure below shows both electricity (elegas=E) and natural gas (elegas=N) seasonal trends for one building during the period considered in this study (January 2018 through December 2019). Electricity use fluctuates considerably over a two-year period given the need for cooling during the summer, heating during the winter and other electricity loads present in the building; while natural gas use remains flat at zero, except for the winter period when the fuel is used for space heating. Based on the general predominance of electricity usage in buildings compared to natural gas, we focused on modeling electricity data only.

FIGURE 11 - Single Building Seasonal Electricity and Natural Gas Usage, 2018-2019

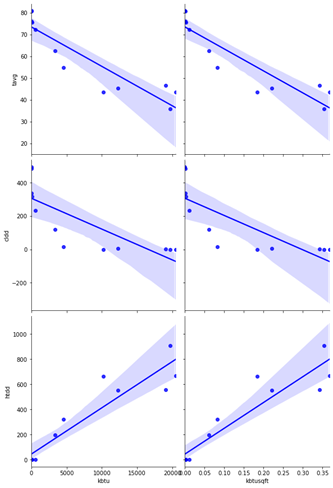
****

To account for cyclic patterns, we used a cyclic encoder to model the months of a year as points on a yearly circle. By using the x,y locations on the circle as inputs, the model learns times of the year that have high energy usage and those with lower energy usage, so that it doesn’t mistake seasonal energy usage trends.

We first tried the Scikit-Learn Linear Regression model. We started with the following building characteristics: ward, year built, and building type. We also used monthly average wind speed (awnd), cooling degree days (cldd), heating degree days (htdd), average monthly temperature (tavg), and total monthly snowfall in millimeters (snow).

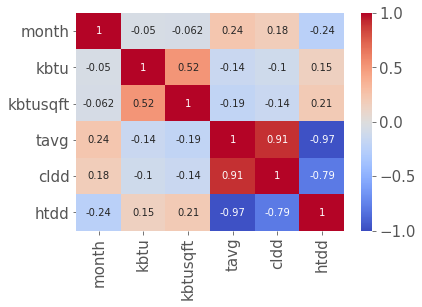
We began examining energy usage in kbtu and energy intensity in kbtu/sqft, as the dependent variable, and weather data and building characteristics as the independent variables.

FIGURE 12 - Regression Plots - Kbtu and Weather Features for a Single Building



The regressions plots above, and heatmap below, show conclusive correlation between energy usage in kbtu and cooling degree days (cldd) for one building’s electricity data; this validates the use of electricity as the energy type that cools the building. On the other hand, a negative correlation between energy usage and heating degree days (htdd) suggests the use of another fuel to heat the building (e.g. natural gas, district steam, etc.). Both are to be expected and follow energy usage slopes as depicted in Figure 1 (page 1).

FIGURE 13 - Heatmap for Single Building Instance



We also noticed a high correlation between average outdoor temperature and cooling degree days (0.91), as well as between average outdoor temperature and heating degree days (-0.97). This is expected as the temperature is used to calculate both heating and cooling degree days.

As a result, we decided to drop tavg as a feature.

We approached the study as a supervised machine learning problem and regressed energy usage against weather conditions and building characteristics at an instance in time, where instance of time is a month. We attempted two separate approaches: 1) building a global model that included all 19 building types, and; 2) modeling each building separately. Then we compared the scores of all models, a model trained on all building types, and models trained on individual building types to identify the best performing models.

We made some decisions on target variables and some features, based on modeling: One Hot encoded Ward and Year Built; used reported area as a feature and energy usage as the target variable.

We built an extraction pipeline using the Scikit-Learn pipeline module. The pipeline includes a ColumnTransformer to combine several feature extraction mechanisms into a single transformer; FeatureUnion to concatenate results of time component transformers; and OneHotEncoder for the three categorical features Ward, YearBuilt and BuildingType.

Times Series Train/Test Split: Since training and evaluating machine learning models usually requires a training set and a test set, we first tried using Scikit-Learn’s train\_test\_split method taking 20% of the data as test data. As weather data changed throughout time, we used time-based cross validation to form a type of “sliding window” training approach using 12 splits. This way, we performed time series cross validation on the model, returning the cross validated Determination of Coefficient (R2), mean squared error (mse), and mean absolute error of the regressor, along with the final fitted model, fitted on all of the data.

Global Model: Power Rangers wanted to observe which model produced better R2 outputs while using different target variable options, as follows:

TABLE 1 - R2 with 1) kbtu, with sqft not included as a feature, 2) energy intensity in kbtu/sqft and, 3) kbtu with sqft included as a feature

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target | Sqft as a feature | R2 using train\_test\_split (ts=0.2) | Mean R2 using TSS | R2 standard deviation using TSS |
| Kbtu | No | 0.174 | 0.19 | 0.03 |
| Intensity (kbtu/sqft) | - | 0.361 | 0.27 | 0.15 |
| Kbtu | Yes | 0.704 | 0.72 | 0.04 |

We can see the effect of each case separately, where standard linear regression does not work very well when excluding square footage from the analysis. For the last case (target = kbtu and sqft as a feature), the R2 scores are consistent across the 12 splits. The mse shows some level of variability.

TABLE 2 - Standard Linear Regression with Target as Kbtu and Sqft as a Feature

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Split | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | AVG |
| R2 | 0.64 | 0.72 | 0.75 | 0.72 | 0.75 | 0.68 | 0.74 | 0.76 | 0.74 | 0.74 | 0.65 | 0.74 | 0.72 |
| mse | 4.43e+11 | 5.71e+11 | 7.17e+11 | 7.10e+11 | 9.79e+11 | 9.41e+11 | 1.04e+12 | 8.19e+11 | 9.47e+11 | 7.66  e+11 | 1.01  e+12 | 5.6  e+11 | 8e+11 |

Once we agreed that energy usage in kbtu would be our target variable, and that square footage (“reportedbuildinggrosssquarefloorarea”) would be included as a feature, we tried other regression models by looping different regressors into our pipeline. Results follow below.

TABLE 3 - R2/MSE Global Modelling All Regression Models

|  |  |  |
| --- | --- | --- |
| Model | Coefficient of Determination | Mean Squared Error |
| SVR | -0.047 | 2.25e+12 |
| Lasso | 0.694 | 8.53e+11 |
| Ridge | 0.703 | 6.75e+11 |
| LinearSVR | -0.004 | 1.02e+12 |
| ElasticNet | 0.731 | 1.05e+12 |
| MLPRegressor | 0.689 | 9.39e+11 |
| KNeighborsRegressor | 0.948 | 1.12e+11 |
| DecisionTreeRegressor | 0.932 | 1.48e+11 |
| RandomForest Regressor | 0.974 | 8.71e+10 |

As the table shows, for the global model (that has building type encoded as a feature), the scores are similar between simple Linear Regression, Lasso and Ridge (around 0.7). What that means is that there is error due to bias because we are using a simpler model that has more noise. It can't discriminate the noise between all these different building types; in the high dimensional space they are all very close together and are not separable.

We also tried a combination of LinearRegression and PolynomialFeatures of second order to try and fit a paraboloid to the data instead of a plane, and also tried third order polynomial features. Results are shown below:

TABLE 4 - LinearRegression and PolynomialFeatures results - global model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean R2 | Std. Dev. R2 | Mean RMSE | Std. Dev RMSE |
| LR-PolyFeat(2) | 0.84 | 0.12 | 681 | 614 |
| LR-PolyFeat(3) | 0.65 | 0.23 | 992 | 811 |

Although the quadratic polynomial model performs better than the linear regression model, it still has a large RMSE, compared to the average energy usage, indicated in the table below:

TABLE 5 - Average Usage (MBtu) - All Buildings:

|  |  |  |
| --- | --- | --- |
| **P25** | **P50** | **P75** |
| 174 | 400 | 900 |

Next, we decided to assess if models based on individual building types might perform better. To that end, we wanted to create a more complex model using an ensemble model: voting regressor, linear regression and random forest regressor.

Individual Building Models: Although we attempted to model each building type, we were unsuccessful with those with smaller instances, eg. tech facilities, utilities and manufacturing/industrial plants, as well as parking structures since modeling unconditioned space did not make sense. Below we share the results of the dataset’s largest building type groups, lodging/residential (49% of all instances) and office (31% of all instances).

TABLE 6 - R2/MSE Lodging/Residential, Ensemble Model

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Split | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | AVG |
| R2 | 0.78 | 0.83 | 0.82 | 0.76 | 0.86 | 0.84 | 0.84 | 0.85 | 0.87 | 0.89 | 0.83 | 0.88 | 0.84 |
| MSE | 7.60e+10 | 5.66 e+10 | 5.84 e+10 | 1.08 e+11 | 5.94 e+10 | 9.58 e+10 | 9.06 e+10 | 7.85 e+10 | 5.80 e+10 | 4.21 e+10 | 5.68 e+10 | 3.52 e+10 | 7e+10 |

TABLE 7 - R2/MSE Office, Ensemble Model

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Split | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | AVG |
| R2 | 0.85 | 0.87 | 0.91 | 0.89 | 0.89 | 0.91 | 0.91 | 0.90 | 0.88 | 0.90 | 0.89 | 0.87 | 0.89 |
| MSE | 1.42 e+11 | 1.23 e+11 | 8.56 e+10 | 9.65 e+10 | 1.02 e+11 | 1.01 e+11 | 1.06 e+11 | 1.21 e+11 | 1.17 e+11 | 9.67 e+10 | 9.55 e+10 | 1.29  e+11 | 1e+11 |

The ensemble model for these building types performs better because it has reduced error due to variance and the increased complexity promotes improved separability.

**Hyperparameter Tuning**

Hyper parameter tuning was attempted on our RandomForestRegressor and VotingEnsemble. Both GridSearchCV and RandomSearchCV were used. The GridSearchCV was computationally expensive to run for both models. Each model was allowed to run for 5 hours before the kernel was interrupted. This led us to attempt a RandomSearchCV. This was less taxing computational but it still took too long to run so again the kernel was interrupted.

**Conclusions and Way Forward**

Owner reported (versus tax reported) square footage is the most important driver of energy usage in buildings larger that 50,000 square feet in DC, as compared to other features, eg. weather and other building characteristics.

Global model’s RMSE is considerably high compared to buildings energy usage interquartile ranges. Linear Regression model with second order polynomial features performs better than third order polynomial.

In addition, the individual building type model performs better than the global model for lodging/residential and office buildings because ensemble models such as Random Forest enhance weaker models such as the global Linear Regression model. We believe our ensemble model will accurately predict energy usage for most building types in DC, but will do an excellent job predicting energy usage for certain building types such as lodging/residential and offices. Other, lower instance, building types that performed well included mixed use property and religious/worship, however, some of scores were so high as to suggest model overfitting.

In the future, as DC’s data matures, we believe the DOEE can better understand how large commercial buildings consume energy using these types of models, perhaps varying features, eg. including rainfall in weather data, expanding on building characteristics such as roof and window types and areas.. With additional ES scores, DC should be able to identify how energy efficient capital improvements reduce energy consumption and over time, adjust building codes to ensure energy efficient building characteristics are included in construction and retrofits.

Given additional time to focus on this problem, Team Power Rangers next steps would include: developing an application to provide predictions given our features of weather and building characteristics; conducting hypertuning to further improve our modeling efforts; looking further at the effects of building location on energy usage; search for datasets with smaller time intervals (daily) to understand the daily cycles of building energy usage, obtain data from calendar year 2020 to understand the effects of COVID-19 on energy consumption profiles, and most important, Team Power Rangers would partner with the DC Government to produce usable model output for DC energy managers.

**Registry/Code Location**

Team Power Rangers code may be found at: https://github.com/georgetown-analytics/Power-Rangers

1. https://ecommons.udayton.edu/phy\_fac\_pub/13/ [↑](#footnote-ref-0)
2. https://portfoliomanager.zendesk.com/hc/en-us/articles/211697117-What-is-an-ENERGY-STAR-score [↑](#footnote-ref-1)
3. https://opendata.dc.gov/datasets/building-energy-benchmarks?geometry=149.414%2C-81.268%2C-42.539%2C69.748 [↑](#footnote-ref-2)
4. <https://www.ncdc.noaa.gov> [↑](#footnote-ref-3)
5. Minutes fromMeeting with DC DOEE, 2 Oct 20, https://drive.google.com/file/d/1aD8bAcxANEjw9emj-PXP-CH\_PO7Uz8Hr/view?ths=true [↑](#footnote-ref-4)
6. https://dc.curbed.com/maps/dc-council-eight-wards-map-elections-neighborhoods [↑](#footnote-ref-5)
7. https://github.com/georgetown-analytics/Power-Rangers/blob/master/Dont\_Look/Map\_Dom.ipynb [↑](#footnote-ref-6)
8. Model Selection Management Systems: The Next Frontier of Advanced Analytics, http://pages.cs.wisc.edu/~arun/vision/SIGMODRecord15.pdf [↑](#footnote-ref-7)
9. DC Building Energy Benchmarks Metadata,https://www.arcgis.com/sharing/rest/content/items/aba010cff7fe4d4cb369a54b56cd7544/info/metadata/metadata.xml?format=default&output=html [↑](#footnote-ref-8)
10. https://link.springer.com/article/10.1186/s40327-018-0064-7 [↑](#footnote-ref-9)