Analyzing Energy Usage in Remote Communities

Data Scientists:

Audrey Jameson, Cristina Ortuño Mojica, Farinaz Kouhrangiha, Jaret André and Meilin Lyu

Mentors:

Aditi Maheshwari, Galen Richardson, Ryan Kilpatrick and Shreyas Choudhary

Collaboratory Notebook

It then splits the data and runs our Random Forest model, displaying detailed accuracy metrics and a confusion matrix. The model is improved by means of feature selection and hyperparameter tuning and a final set of classifications is appended onto the original .csv file.

Introduction

Remote communities are long-term settlements with at least 10 dwellings which are not connected to the North American electrical grid or natural gas pipeline network. As a result of the extreme location and climate of northern remote communities in Canada, energy costs are high and the deterioration of buildings and infrastructure are unavoidable. Additionally, since these communities are off the electricity grid, diesel accounts for a majority of their energy usage. Before we can make the transition to more sustainable forms of energy, however, it is crucial that we gain a better understanding of the current energy demand. Given that energy usage in these communities is largely allocated to heating, our challenge was to predict the heating load of each dwelling in Cambridge Bay, NT and in part, identify the reasonable expectation of renewable energy for any remote community going forward.

Our objective was to build a classification model to predict building types as commercial, residential, or industrial, with respect to QGIS building features such as area, perimeter, point count, building height, and volume, among others. With these classifications and other characteristics, as well as more domain knowledge, the heating load for each building can then be estimated.

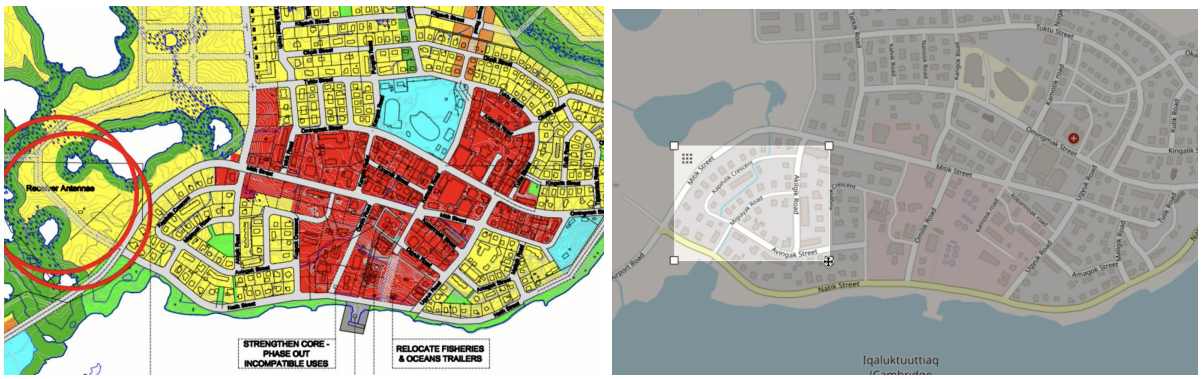
Data Collection

We started this project by collecting information about Cambridge Bay, a remote community in Nunavut which was to be our focus for the project. Due to the uniqueness of the buildings in remote communities, we were unable to find a pre-existing dataset of similar buildings which we could use to train our model. It then became our task to extract characteristics of the buildings in Cambridge Bay ourselves using geographic information system (GIS) data. Using a free GIS application called QGIS, we were able to visualize, manipulate and analyze GIS data on the community. Extracting the building polygons from these GIS datasets was a large learning curve since GIS is often considered its own sector and none of us had experience with it prior. We also had to account for heavy distortions from projections as a result of Cambridge Bay being a northern community.

Next, we followed the *Wurm 2020* paper to feature engineer several new complex geometric building characteristics that allowed us to get more information on the buildings which the machine learning model could use to learn more about the distinct attributes for the different building classifications. We also used DSM, DTM, and DEM files from an open source to collect 3D GIS data which gave us insights on the building heights. These datasets were very heavy in terms of memory and the computational time needed to process them as a single 25km x 25km square was roughly 2GB and extracting the 2D and 3D characteristics took about an hour per community on an average computer.

Before we could move forward with building our classification model, it was also necessary to complete our dataset by labelling our collected data with the appropriate classification. Our initial approach was to manually label each building using OpenStreetMap and Google Street View, as well as any other information we could find on the community's infrastructure. This was ultimately tedious, error-prone, and certainly not a sustainable way to build out the dataset for future use. We decided to switch tactics to make the labelling process less labour intensive while simultaneously diversifying our dataset and allowing for quick and easy expansion.

Inspired by methods used in the *Wurm 2020* paper and using community planning maps of other remote communities in Nunavut as reference, we were able to extract small plots of land designated residential, commercial, and industrial from OpenStreetMap, export the .osm files to be inputted to a QGIS model, and from there uniformly assign labels to all the buildings within the plot. This allowed us to quickly expand our dataset to include over 1000 buildings from 11 different remote communities in Nunavut instead of the 1 we had originally set out to label through a process that could easily be replicated to further expand the dataset. While inaccuracies were inevitable given exceptions to land designation zonings, our options were limited given our timeline and need for data.

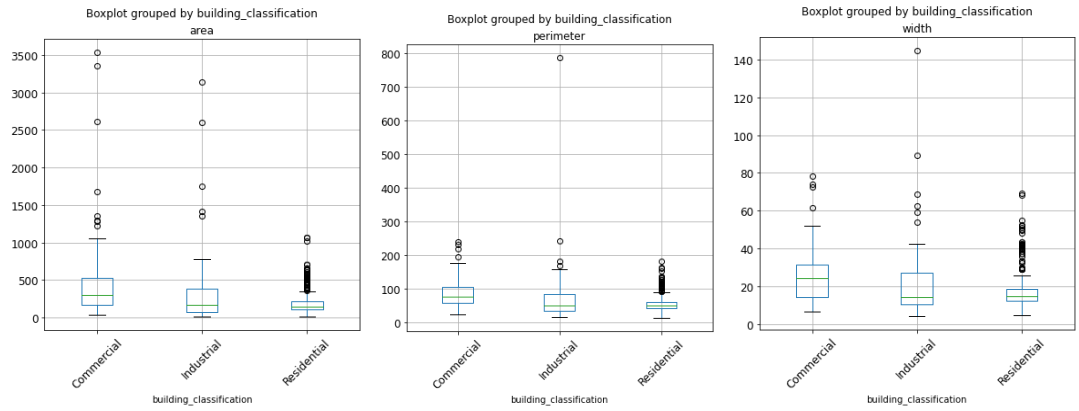


Community planning maps (left) were used to extract subplots of land having the same land designation from OpenStreetMaps (right).

Exploratory Data Analysis

Before moving forward with our classification model, we first performed some exploratory data analysis on our current dataset which includes over 1000 buildings distributed over Cambridge Bay as well as other communities in Nunavut.

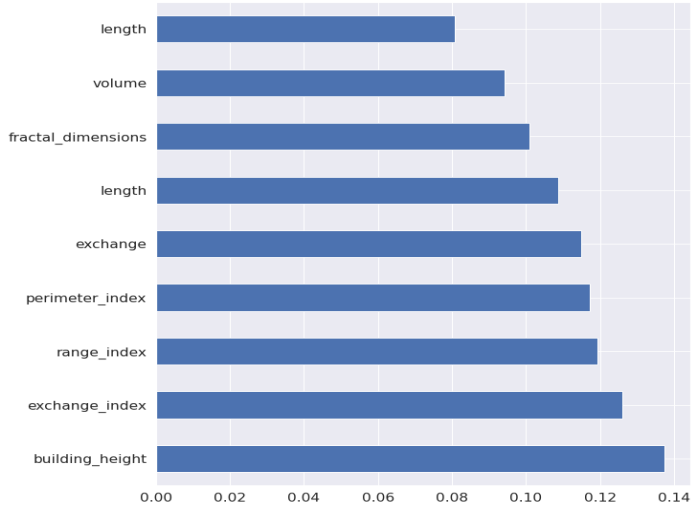
Some notable trends we found were a low variation in residential building area when compared to industrial and commercial as well as similar trends with regards to height, perimeter, and width which were to be expected.



ML Model

After preparing the dataset and normalizing it, we began building our classification model. We chose to use a random forest model for the task and utilized the Python scikit-learn library. For our purposes, the data was split into a test set consisting of all dwellings in Cambridge Bay and a training set of the remaining data. This training set was then balanced through oversampling to compensate for the high number of residential buildings in the dataset.

In order to increase the accuracy and obtain a model with optimal performance, we choose to use the Embedded method, L1 Regularization and univariate selection algorithms for feature selection. The embedded method is relatively faster as a filter method and obtains a good accuracy. L1 Regularization works well for large datasets because it treats less important features as coefficient 0, while the univariate method selects features based on statistical tests. All of the three methods recorded the same results for the importance of features, as shown below.



The feature importance results from the Embedded method.

We also tuned our hyperparameters such as test size, max depth, random state, cv, n-estimators, min-samples-split, n-jobs, n-split, n- repeat etc. using a random hyperparameter grid search and grid search with cross validation. These modifications improved our model slightly.

Results

The project deliverable expected for this project was inference, visualizations and a theoretical process using ML techniques to estimate the heating consumption for one community with the potential to be expanded to other remote communities. We decided to adapt this deliverable to a reproducible pipeline to collect, aggregate, and cleanse a dataset that will then run a model and produce several outputs including the classification of all buildings in a community. We managed to curate a diversified dataset and build a classification model which can now be used to produce heating estimates and puts us one step closer to transitioning these remote communities in Canada away from diesel. As for the metrics of our model itself, we achieved ~65% accuracy, however in order to grant any validity to the results and metrics of a model, it's crucial that the dataset be good, balanced, and unbiased. Due to our conventional methods of data collection and limited resources available, we don't believe this number necessarily reflects the quality of our model but rather the quality of our data.

Next Steps

The next steps for this project would be to expand our existing dataset and improve our classification model utilizing semi-supervised learning methods and other ML techniques. Specifically, given the natural clustering of infrastructure we believe our model would perform better if it took into account a building's surroundings during the classification process.

Another future step would be using the building archetypes since not all buildings in a building classification would have the same calculation for heating consumption. For example, garages and single-unit residentials would both be classified as residential but garages would use significantly less heat since the average garage would not be heated 24/7.

In order to produce accurate and meaningful heating estimates, more research also would have to be done on the typical energy usage of specific building types. After all, if the heating load were only to be calculated as a function of volume, there would have been no purpose for our classification model.

Conclusion

While our finished project was not exactly what we had expected and continued to evolve throughout the 8-week process, in many ways it went beyond our goals. We all learned and grew in various ways with the challenges we were faced with and having the opportunity to work on a real world problem was an invaluable experience. There were definitely learning curves which were to be expected when approaching a problem involving new tools and technologies, but working through it collaboratively with a group and learning from each other made the process very enjoyable. Exploring a topic we may not otherwise have been exposed to was also beneficial to us and we hope our efforts will show the potential in a larger-scale initiative to aid in the transition away from diesel in these remote communities.

Resources

1. http://www.buildingnunavut.com/en/communityprofiles/communityprofiles.asp
2. https://www.mdpi.com/2220-9964/10/1/23
3. https://www.pembina.org/pub/diesel-reduction-progress-remote-communities
4. http://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm
5. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E
6. https://www.mdpi.com/2220-9964/10/1/23/htm
7. https://scikit-learn.org/stable/modules/feature_selection.html

8. https://machinelearningmastery.com/random-forest-ensemble-in-python/
9. https://www.datacamp.com/community/tutorials/random-forests-classifier-python
10. https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c
11. https://stackoverflow.com/
12. https://programmer.group/
13. https://towardsdatascience.com/
14. https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/
15. https://machinelearningmastery.com/tune-number-size-decision-trees-xgboost-python/
16. https://www.hackerrank.com