



Project Overview

Project Context

Role:

- You should act as a consultant for eSC, an energy company

Goal:

- How might your client predict future energy usage?
 - Predict energy usage if the summer was 5 degrees warmer
 - Provide actionable insight into how to reduce energy costs

What data to analyze – should be:

- A function of what the team determines might be useful
- Determined by each project team
- There is *A LOT* of data

Remember this needs to be data driven –

Project Data

Static House Data

A file with basic house information for a random sample of single-family houses that eSC serves.

- The file contains the list of all houses in the dataset.
- For each house, there is information describing the house.
 - The information ranges from the building id (used to access the energy data) to other house attributes that do not change (such as the size of the house).
- There are >5,000 houses in the dataset (rows in the file)
- The file can be found at:
https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/static_house_info.parquet

Note that this file is in 'parquet' format an optimized binary format

Energy Usage Data

- For each house, there is a file that contains energy usage data, which was collected hour-by-hour.
- There is one data file per house. Energy usage is:
 - Collected every hour
 - Collected across many sources (ex. air conditioning system, dryer
 - the 'building ID' is file name which identifies the house.
- Note that each file is in 'parquet' format.
- All the data is in one folder on Amazon AWS.
- For example, the following URL is for 'building_id' 102063.
<https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-houseData/102063.parquet>

There are more than 5,000 houses (i.e., different building id's) in the directory

Weather Data

- Hour-by-hour weather information (one file for each geographic area)
- The weather data was collected for each county and stored based on a county code:
 - The county code for each house can be found at 'in.county' column of the house static dataset. This file is in a simple CSV format.
 - For example, the following URL provides the weather for county 'G4500010'.
<https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/weather/2023-weather-data/G4500010.csv>

There are approximately 50 counties in the directory.

Meta Data

- A data description file, explaining the fields used across the different housing data files.

~270 attributes

https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/data_dictionary.csv

Project Approach

Suggested Project Steps

- a) Determine the best approach to read and merge the data.
- b) Determine what should be the output during this 'data preparation' phase.
- b) Do exploratory analysis of the data.
- c) Build a model that predicts the energy usage, for any given hour, for the month of July.
 - July was selected, as eSC thought July is typically the highest energy usage month.
 - Hint: you will need to try several models and pick the best model.
- d) Understand and be able to explain your model's accuracy.

Suggested Project Steps

- e) Create a new weather dataset → all July temps 5 degrees warmer.
- f) Use your best model to evaluate peak future energy demand.
 - assume no new customers
 - Note: this must be model driven, not just increasing energy usage by a percentage
- g) Show future peak energy demand in total (for an hour):
 - For different geographic regions
 - For other dimensions /attributes you think important
- h) Create a shiny application to interact with the data.
- i) Identify one approach to reduce peak energy demand.
- j) What would you suggest, how would you model the impact?
- k) How would you explain the impact? BE DATA DRIVEN!

Project Deliverables

Project Deliverables

Word Document:

- Target audience is your manager / instructor
(hint: your manager/instructor is a data science expert)
- Focus on what was accomplished
- Should describe all analysis done, even if an analysis did not generate any interesting results, it should still be included

Presentation:

- Target audience is your client (*hint: the client is not a data science expert*)
- Presentation length is 10 minutes (*lab instructor will explain specifics*)
- Be sure to include the following in your presentation:
 - Number of records in dataset evaluated
 - Key drivers identified; accuracy of results

Project Deliverables (continued)

Interactive Application (shiny app):

- a) A shiny app needs to be created and deployed on shinyapps.io
- b) To better understand your model's energy prediction
- c) To better understand the potential future energy needs and/or savings

Expectations

- 1) Work at a consistent pace throughout the rest of the semester
- 2) Tasks should be distributed equally across the team members
- 3) Tasks should typically not take a long time to complete – one week target, two weeks is fine, but not a month
- 4) Tasks should be at an appropriate level of effort / detail

Project Updates

- 1) Project Updates: Nov 17, Nov 24, Dec 1
(one per group, not per person)
- 2) For each update (including for the final submission), provide:
 - a) Work done by each person (since the last update)
 - b) Work planned to be done by each person (by the next update)
 - c) Key issues / challenges



Project Grading

Presentation (5%)

0.5% - Business Questions – Describe the goals of the project

0.5% - Use of Descriptive statistics - Provide context and a basic understanding of the data

1% - Use of modeling techniques – Show the results of different models and explain why they were/were not useful

1% - Visualization - Convey the results in an easy-to-understand manner

1% - Interpretation of the results/Actionable Insights – Make sure the results are actionable (as compared to just interesting)

1% - Know your audience - Present findings in an easy-to-understand way (ex. no data science lingo, easy for others to follow the logic)

Shiny App (5%)

1% - App can load / use a data file provided by the user

0.5% - Display of the first 'n' rows of the read in dataset

1% - Generate predictions via a stored model

0.5 % - Display the Confusion Matrix

2% - An explanation within the app, of how to interpret the Confusion matrix:

- Which numbers to “look at”
- What is a “good number”

Word Document (15%)

- 1% - Business Questions** - Describe the goals of the project
- 1% - Data cleanse/munge/preparation** - Transform/clean/munge the data appropriately, including missing values
- 1% - Use of Descriptive statistics** - Provide context and a basic understanding of the data
- 4% - Use of modeling techniques** - Try at least 3 different models and compare results
- 3% - Visualization** - Convey information in an easy-to-understand manner
- 4% - Interpretation of the results/Actionable Insights** – Make sure the results are actionable (as compared to just interesting)
- 1% - Validation** - How do you know your results were correct (i.e., no errors)?

Final Project (Word doc)

Example Table of Contents:

- Introduction (scope/context/background)
- Business Questions addressed
- Data Acquisition, Cleansing, Transformation, Munging
- Descriptive statistics & Visualizations
- Use of modeling techniques & Visualizations
(noting techniques explored but not used in presentation)
- Actionable Insights / Overall interpretation of results