Exam PA October 2024 Project Statement

IMPORTANT NOTICE – THIS IS THE OCTOBER 25, 2024, PROJECT STATEMENT. IF TODAY IS NOT OCTOBER 25, 2024, SEE YOUR TEST CENTER ADMINISTRATOR IMMEDIATELY.

General Information for Candidates

This examination has 10 tasks numbered 1 through 10 with a total of 70 points. The points for each task are indicated at the beginning of the task, and the points for subtasks are shown with each subtask.

Each task pertains to the business problem described below. Additional information on the business problem may be included in specific tasks—where additional information is provided, including variations in the target variable, it applies only to that task and not to other tasks. This exam includes an Excel data file with information for Task 5. You may use Excel for calculation for any of the tasks, but all answers must be submitted in the Word document. *If you upload the Excel document, it will not be looked at by the graders.* Neither R nor RStudio are available.

The responses to each specific subtask should be written after the subtask and the answer label, which is typically ANSWER, in this Word document. Each subtask will be graded individually, so be sure any work that addresses a given subtask is done in the space provided for that subtask. Some subtasks have multiple labels for answers where multiple items are asked for—each answer label should have an answer after it.

Each task will be graded on the quality of your thought process (as documented in your submission), conclusions, and quality of the presentation. The answer should be confined to the question as set. No response to any task needs to be written as a formal report. Unless a subtask specifies otherwise, the audience for the responses is the examination grading team and technical language can be used.

Prior to uploading your Word file, it should be saved and renamed with your five-digit candidate number in the file name. If any part of your exam was answered in French, also include “French” in the file name. Please keep the exam date as part of the file name.

The Word file that contains your answers must be uploaded before the five-minute upload period time expires.

Business Problem

You are working for a firm that consults on energy use in the United States. Your firm serves a wide range of clients including energy producers, municipal governments, developers, and building owners. Your clients are interested in using data to understand patterns of existing energy use and to predict energy usage in the future.

Your firm is currently focused on the Chicago market and will use data from the city of Chicago[1] that looks at detailed energy use at the census block[2]-level in residential, commercial, and industrial buildings. The energy data has also been enriched with weather data.[3]

Notes on the data set:

Energy data is collected at the census block level for specific building types (residential, commercial, and industrial).

U.S. census data is organized at different levels of granularity with the census block being the most granular. A census block within an urban area typically represents a single city block.

The weather data is captured at a daily level and is collected from Chicago's Midway airport and used for the entire city.

[1] *Source: City of Chicago – Chicago Data Portal*

[2] A census block is the smallest area used by the U.S. Census Bureau. In a city, it is typically an area bounded by four streets with no streets running through it. They will be referred to as simply "blocks" in this assessment.

[3] *Source: National Oceanic and Atmospheric Administration – National Centers for Environmental Information*

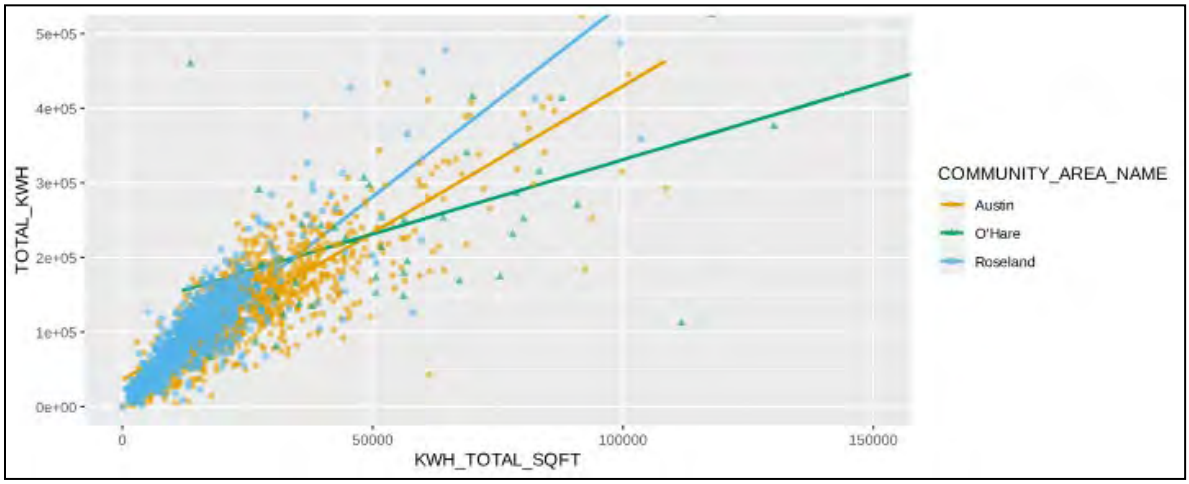
Data Dictionary

Variable	Data Type / Range / Example	Description
COMMUNITY_AREA_NAME	Character: Roseland, O'Hare, Austin, etc.	Name of the neighborhood in Chicago. Each neighborhood comprises many census blocks.
BLOCK_HOUSING_OCCUPIED_UNITS	Numeric	Number of occupied housing units on the block
KWH_TOTAL_SQFT	Numeric: 600 - 5453937	Total square footage for a census block and building type
BUILDING_AGE or AVG_BLDG_AGE	Numeric: 0 - 158	Average age of the buildings in a census block
STORIES	Numeric: 1 - 110	Average number of stories for buildings in a census block
BUILDING_TYPE	Character: Commercial, Residential, Industrial	The type of building based on what it is used for, possible values are Commercial, Residential, and Industrial
AVERAGE_HOUSEHOLD_SIZE	Numeric: 1.28-4.37	Average household size. Measured at the census tract level.
OVER_AGE_65	Numeric: 0.23-0.45	Proportion of the population in the census tract over age 65.
GAS_ACCOUNT	Numeric: 1 - 383	The number of different gas accounts associated with a census block
ELECTRICITY_ACCOUNTS	Numeric: 1 - 1904	The number of different electricity account associated with a census block
KWH_JAN, KWH_FEB, …	Numeric: Varies	Total KWH usage for that census block and building type for a month. 12 variables, one for each month. KWH is a unit of measurement for electricity.
TOTAL_KWH	Numeric: 312 - 72,070,053	Total KWH usage for that census block and building type for a year.
THERM_JAN, THERM_FEB, …	Numeric: Varies	Total therm usage for that census block and building type for a month. 12 variables, one for each month. A therm is a unit of measurement for natural gas.
TOTAL_THERM	Numeric: 28 - 1,940,742	Total therm usage for that census block and building type for a year.
THERMS_PER_SQFT	Numeric	Total therm usage divided by total square footage for a census block and building type.

KWH_PER_SQFT	Numeric	Total KWH usage divided by total square footage for a census block and building type.
THERMS_PER_ACCOUNT	Numeric: 0 - 43,222	Total therm usage divided by the number of gas accounts.
TMAX_FAHRENHEIT	Numeric: 12 - 95	Max temperature recorded during the day in degrees Fahrenheit.
TMIN_FAHRENHEIT	Numeric: 0 - 79	Min temperature recorded during the day in degrees Fahrenheit.
PRECIPITATION_INCHES	Numeric: 0 - 4.77	Inches of precipitation during the day.
SNOW_FALL_INCHES	Numeric: 0 - 8.8	New snowfall during the day.
SNOW_DEPTH_INCHES	Numeric: 0 - 9	Snow depth reported as 7 am each day.

Task 1 – (6 *points*)

Your client wants to understand the relationship between electricity consumption and total square footage in residential buildings from different community areas, in particular Austin, Roseland and O'Hare. Your assistant created the graph below filtering only residential building energy use across the three community areas. The lines represent ordinary linear fit within each community area. Your manager suggests that adding an interaction between **COMMUNITY_AREA_NAME** and **KWH_TOTAL_SQFT** is necessary to capture the slope differences.



(a) (2 points) Assess your manager's suggestion that adding this interaction can capture the slope differences.

Candidates performed very well on this task, with most candidates receiving full credit. Full-credit answers took a stance on whether the recommendation is valid, demonstrating an understanding of how interaction terms work. Strong recommendations against capturing an interaction term because of sparsity of O'Hare datapoints were awarded full credit.

ANSWER:

I agree with the manager's suggestion because it allows each community to have its own slope coefficient.

The graph shows different slopes for KWH_TOTAL_SQFT on TOTAL_KWH across Austin, Roseland, and O'Hare, indicating that the relationship varies by community area. Without the interaction term, the model assumes a uniform effect of KWH_TOTAL_SQFT on TOTAL_KWH for all areas.

Your assistant built a GLM with interaction and provided you with the model summary.

```
Call:
glm(formula = TOTAL_KWH ~ KWH_TOTAL_SQFT * COMMUNITY_AREA_NAME,
    family = gaussian(), data = energy_data)

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              3.607e+04  2.542e+03  14.190  < 2e-16 ***
KWH_TOTAL_SQFT                           3.938e+00  8.897e-02  44.259  < 2e-16 ***
COMMUNITY_AREA_NAMEO'Hare                9.530e+04  8.564e+03  11.128  < 2e-16 ***
COMMUNITY_AREA_NAMERoseland             -1.381e+04  3.651e+03  -3.783 0.000159 ***
KWH_TOTAL_SQFT:COMMUNITY_AREA_NAMEO'Hare   -1.941e+00  1.305e-01 -14.874  < 2e-16 ***
KWH_TOTAL_SQFT:COMMUNITY_AREA_NAMERoseland  1.245e+00  1.736e-01   7.176 9.74e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 2010312693)

    Null deviance: 1.3487e+13  on 2225  degrees of freedom
Residual deviance: 4.4629e+12  on 2220  degrees of freedom
AIC: 54009

Number of Fisher Scoring iterations: 2
```

(b) (2 points) Calculate the impact of a 10 square foot increase on total energy usage for each community based on the model above:

a. O'Hare
b. Roseland
c. Austin

Show your work.

Candidate performance was mixed on this task. Common mistakes were incorrectly using (or ignoring) the non-interaction KWH_TOTAL_SQFT coefficient and calculating the prediction for a 10 square foot building.

ANSWER:

Increase in energy usage:

- O'Hare: 3.938 x 10 − 1.941 x 10 = 19.97 KWH
- Roseland: 3.938 x 10 + 1.245 x 10 = 51.83 KWH
- Austin: 3.938 x 10 = 39.38 KWH

(c) (2 points) Recommend and justify one additional variable to include in the modeling that could help explain the variation in electricity consumption per square foot across different communities.

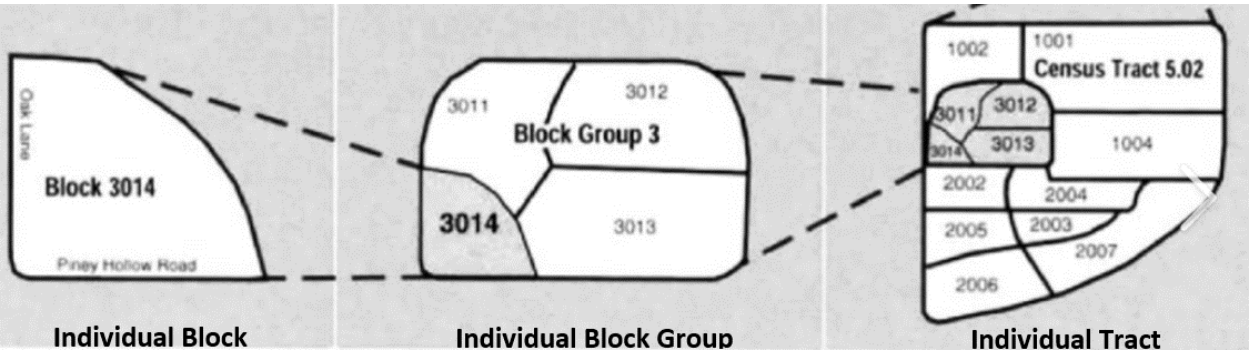
Candidates performed well on this task overall. Any well-justified recommendation was awarded credit. Minimal partial credit was awarded to candidates who provided a valid recommendation without a justification.

ANSWER:

I recommend building age variable (BUILDING_AGE). Older buildings typically have different energy consumption patterns compared to newer buildings due to differences in insulation, construction materials, and energy efficiency technologies.

Task 2 – (6 *points*)

The relationship between a census block and census tract is illustrated below. Census blocks are a geographical subset of a tract.



Your manager wants you to predict residential energy usage in a block for the entire year. Some of the predictor variables are stored at a more aggregate granularity than the block level (e.g., tract and city level detail). Given this, your manager asks you to remove all tract level and city level variables as they will not provide predictive power.

(a) (*2 points*) Critique your manager's recommendation to remove city-level data.

Candidate performance was mixed on this task. Full-credit answers supported their conclusion with a discussion of how city-level variables would impact the model's performance in this context of predicting energy usage at the block level. Some candidate answers incorrectly assumed that the dataset contains data from more than one city.

ANSWER:

I agree with the manager's recommendation. Adding city-level detail will not improve the predictive model since all the data is from the same city. For example, in a linear model, a city level variable would have a single level and therefore no coefficient estimates. This is because the values would be constant for all data points as this data set only spans one city. We should remove city level detail for prediction of block level energy usage in the entire year.

(b) (2 *points*) Critique your manager's recommendation to remove tract level data.

Candidate performance was mixed on this task. Full-credit answers discussed how tract level variables can improve model performance and the limitation that those variables do not vary across blocks within a tract.

ANSWER:

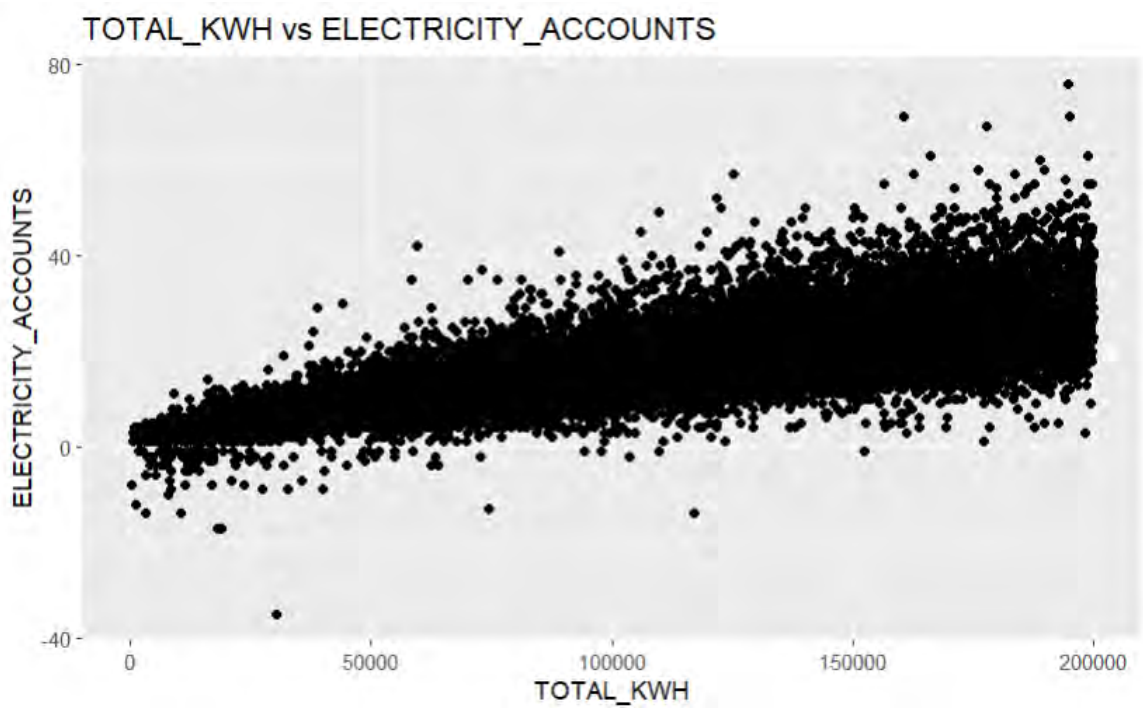
I disagree with the manager's recommendation. Although the tract level detail is less granular, it can still provide predictive power for differentiating block energy consumption which reside in different tracts.

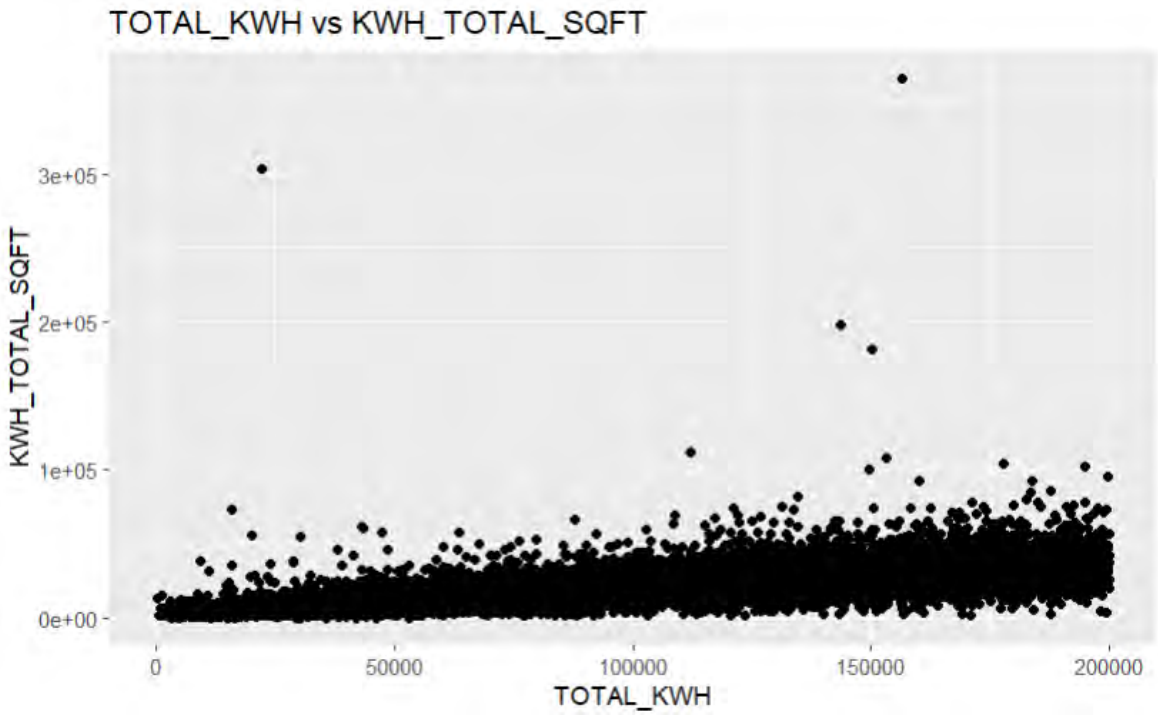
While this will not improve predictive power of block level energy consumption within a single tract, it will improve the overall model as there are many levels to the tract variable within our dataset.

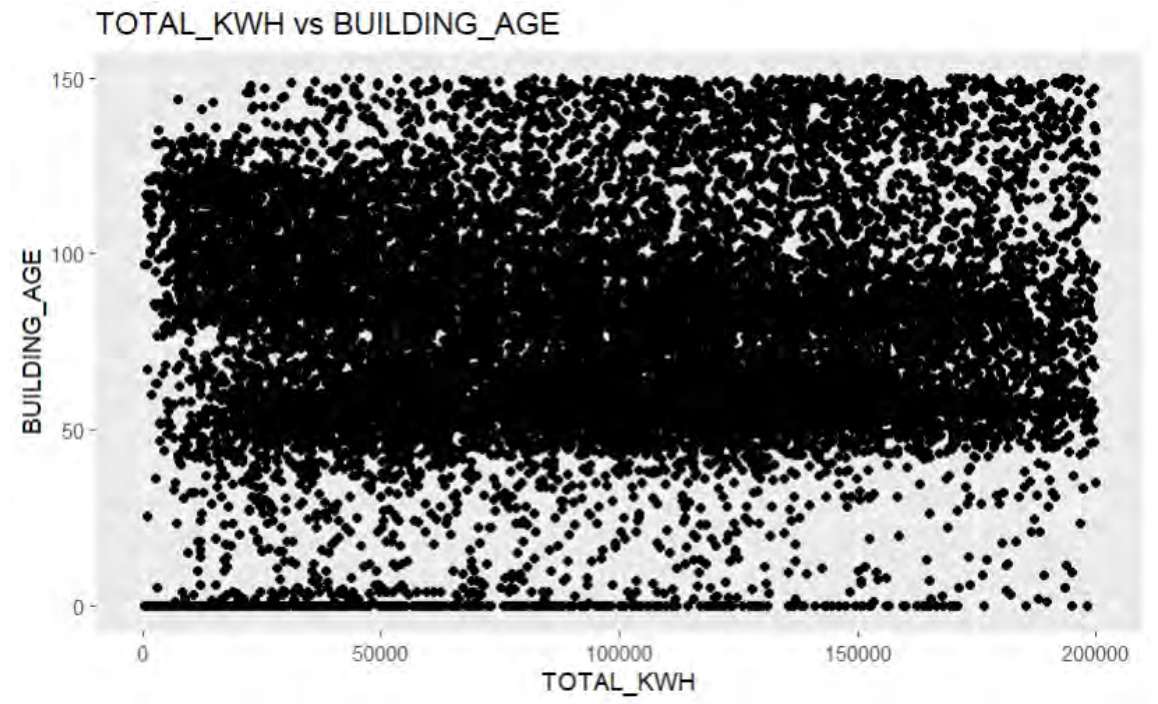
Your assistant wants to model the target variable **TOTAL_KHW** using the formula:

TOTAL_KWH ~ ELECTRICITY_ACCOUNTS + KWH_TOTAL_SQFT + BUILDING_AGE

They have also provided thee charts:







(c) (*2 points*) Recommend whether a tree-based model or a generalized linear model would be a better choice for predicting the target variable. Justify your recommendation.

Candidates performed very well on this task, with most candidates receiving full credit. Full credit was awarded to either recommendation, provided a sound justification based on the charts was provided.

ANSWER:

I recommend using a generalized linear model. Based on the provided charts, the target variable seems to have a linear relationship with at least two of the variables. Given this, a GLM would likely provide a more intuitive model and do a better job modeling the linear relationships. Also, trees tend to perform better with categorical predictor variables. In this case, all three predictors are continuous variables.

Task 3 – (8 *points*)

Your manager would like to identify and predict the highest energy usage blocks. Your assistant created an initial classification model that classifies values as True if they exceed a certain threshold or False otherwise. The output of the model's confusion matrix is shown below.

		ACTUAL	
		False	True
PREDICTION	True	4	660
	False	21493	616

(a) (4 *points*) Calculate the following values from the confusion matrix, and interpret the significance of the results.

a. Accuracy
b. Sensitivity
c. Precision

Candidates performed reasonably well on this task. Correct calculations and strong interpretations of each metric were both required for full credit.

ANSWER:

	Value	**Formula**
accuracy	97.28%	(TP+ TN) / n
precision	99.40%	TP / (TP + FP)
sensitivity	51.72%	TP / (TP + FN)

Accuracy refers to the model's ability to accurately reflect the True Positives and True Negatives, out of the whole prediction space. The model is displaying high accuracy of 97%. At a first glance this implies it is doing a good job at differentiating the True and False values of the data set. However more information is needed.

Precision is the value of True Positives over all predicted positive values. This tells us how well the model is at accurately predicting Positive values compared with false positives. We see a high level of precision of 99%. This indicates the model, when classifying an observation as True, will be very precise in its conclusion and not generate many False Positives.

Sensitivity is the measure of how well the model predicts true positives compared with predicting false negatives. This tells us how sure we can be that the model will capture the true positives in the data set, irrespective of the number of false positives it predicts. We see the sensitivity of the model is much lower than the other metrics at 52%. This indicates that the model will miss a lot of True values and incorrectly classify them as False. There is an inherent tradeoff between increased precision versus increased sensitivity. We need more information to determine if this in an acceptable level of model performance for the sensitivity metric.

Your manager is interested in a marketing promotion targeting high energy utilization homes. Your manager's goal is to identify more of the high utilization homes in the prediction model even at the cost of misclassifying some low energy utilization homes as high energy utilization.

(b) (*2 points*) Recommend how to change the cutoff value of the model to achieve the desired objective. Explain the directional impact of the change.

Candidates performed very well on this task, with most candidates receiving full credit. Full-credit answers recommended reducing the cutoff value and explained how this increases the model's sensitivity.

ANSWER:

I recommend reducing the cutoff value to improve the sensitivity of the model. A classification model starts by predicting for the probability that each observation is True. Secondly, the modeler assigns a cutoff value such that all probabilities above the threshold will be classified as True.

Since we are interested in increasing the sensitivity of the model, lowering the cutoff probability will allow the model to predict more True values and consequently capture more True Positives. This will have a negative impact of also capturing more false positives as well.

Since the sensitivity metric is defined as: True Positives / (True Positives + False Negatives), lowering the cutoff will improve this metric.

Your manager observes that the data set used to train the model is unbalanced.

(c) (*2 points*) Recommend and explain one method to improve model performance by creating a more balanced data set.

Candidates performed very well on this task, with most candidates receiving full credit. In addition to the model solution, alternatives such as undersampling and valid recommendations to use weights to improve model performance were awarded full credit. Partial credit was awarded to answers that discussed relevant techniques without making a recommendation.

ANSWER:

We can create a more balanced training set by utilizing oversampling. This method works by balancing the number of Positives and Negatives in the training data by generating duplicates of the value with fewer observations in the data set. In our specific case, there are significantly fewer True values than False values. So, by oversampling we would duplicate each of the True records until there were as many as there are False records.

Task 4 – (5 *points*)

Your manager would like you to build three tree-based models and tune their hyperparameters. The three models will be a decision tree, random forest, and boosted tree.

(a) (1 point) List one common hyperparameter that could be tuned for all three models.

Candidates performed very well on this task. Identifying a correct hyperparameter without further discussion was sufficient to receive full credit.

ANSWER:

A common hyperparameter that can be used across all three models is "Maximum Tree Depth."

(b) (2 points) Describe a unique parameter for tuning a random forest model and a unique parameter for tuning a boosted tree.

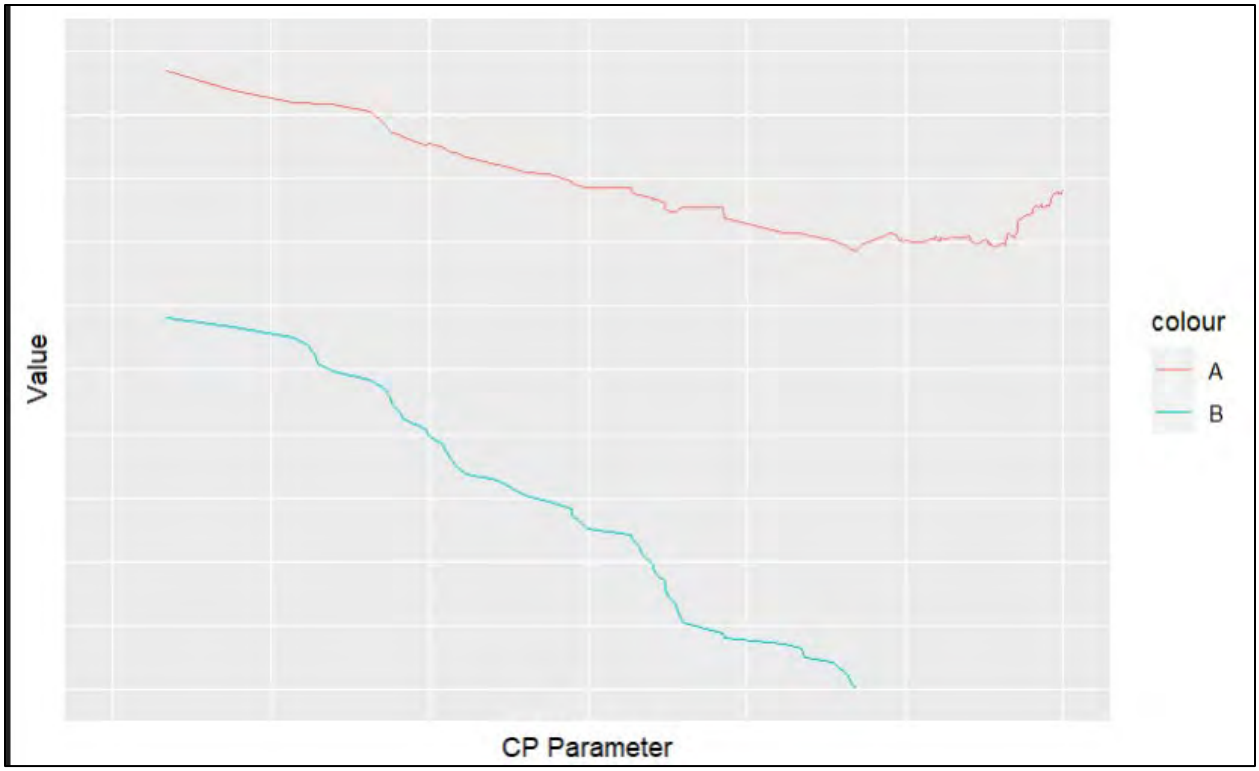
Candidates performed very well on this task, with most candidates receiving full credit. Valid parameters other than the one described in the model solution, e.g. the size of each bootstrap sample, were also awarded full credit, provided they were accurately described. Partial credit was awarded for answers that only listed hyperparameters without providing a description.

ANSWER:

For the random forest we can tune the parameter that sets the number of predictor variables that are sampled for use in each split.

In the boosted tree we can tune the learning rate used as each tree is built, which affects how quickly the loss value is reduced for each subsequent tree.

Your assistant creates a single decision tree and is trying to tune the hyperparameters, specifically the Complexity Parameter (cp). Creating a CP Table, they show the cp values decreasing across the x-axis from left to right and the corresponding values of the error terms for "xerror" and "relerror."



(c) (*2 points*) Identify which line corresponds to the xerror (cross-validation error) versus the relerror (relative error) term and provide a rationale. Explain the behavior of each.

Candidates performed well on this task overall. Full-credit answers recognized the "U" shape to Line A and explained why this is a feature of cross-validation error.

ANSWER:

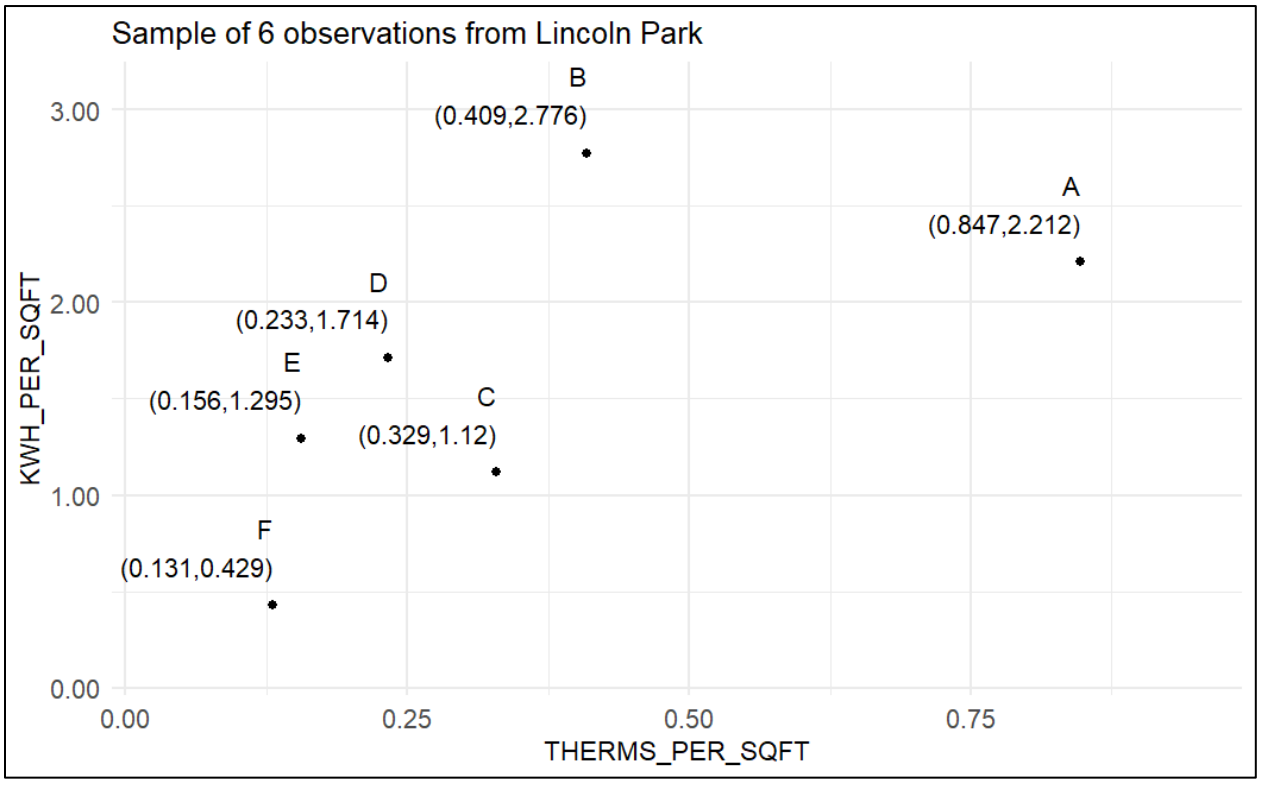
Line A, the top line, corresponds to the xerror, and the bottom Line, B, corresponds to the relerror term. This is because we would expect the relative error to continue decreasing as the cp parameter decreases. However, for the xerror or cross validation error it decreases until a point, then starts increasing as the cp parameter reductions cause the decision tree to overfit.

Task 5 – (7 *points*)

Note: The data for this task is provided in the available Excel file on the "5" tab. You may use it for your calculations, but if the file is uploaded it will not be looked at by the graders. All your work must be shown in this Word document.

Your manager wants to better understand how hierarchical clustering works and asks you to create a dendrogram using the single linkage method on a subset of the energy data. Your assistant selects two variables from the dataset: THERMS_PER_SQFT and KWH_PER_SQFT, scales them, and chooses six scaled observations from the Lincoln Park community area.

	THERMS_ PER_SQFT	KWH_ PER_SQFT
A	0.847	2.212
B	0.409	2.776
C	0.329	1.120
D	0.233	1.714
E	0.156	1.295
F	0.131	0.429



(a) (2 points) Complete the distance matrix below by calculating the Euclidian distances between the missing pairs of observations and enter them into the table below. Round to two decimal places.

Candidates performed very well on this calculation task. The first table below is unchanged from what was provided with the exam, and the second table is completed with the correct values.

ANSWER:

	A	B	C	D	E	F
A	0.00	X	X	X	X	X
B		0.00	X	X	X	X
C	1.21	1.66	0.00	X	X	X
D		1.08	0.60	0.00	X	X
E	1.15	1.50	0.25	0.43	0.00	X
F	1.92	2.36		1.29	0.87	0.00

	A	B	C	D	E	F
A	0.00	X	X	X	X	X
B	0.71	0.00	X	X	X	X
C	1.21	1.66	0.00	X	X	X
D	0.79	1.08	0.60	0.00	X	v
E	1.15	1.50	0.25	0.43	0.00	X
F	1.92	2.36	0.72	1.29	0.87	0.00

Based on the distance matrix, the first cluster formed is with observations C and E.

(b) (1 point) Complete the updated distance matrix using single linkage.

Candidates performed very well on this calculation task. Some candidates incorrectly used values from part (a). The first table below is unchanged from what was provided with the exam, and the second table is completed with the correct values.

ANSWER:

	A	B	C,E	D	F
A	0.00	X	X	X	X
B	0.71	0.00	X	X	X
C,E			0.00	X	X
D	0.79	1.08	0.43	0.00	X
F	1.92	2.36	0.72	1.29	0.00

	A	B	C,E	D	F
A	0.00	X	X	X	X
B	0.71	0.00	X	X	X
C,E	1.15	1.50	0.00	X	X
D	0.79	1.08	0.43	0.00	X
F	1.92	2.36	0.72	1.29	0.00

(c) (2 points) Complete the distance matrices using the tables below to provide the information needed to construct the dendrogram.

Candidates performed well on this calculation task overall. The first three tables below are unchanged from what was provided with the exam, and the following three tables are completed with the correct values.

ANSWER:

	0.00	X	X	X
		0.00	X	X
			0.00	X
				0.00

	0.00	X	X
		0.00	X
			0.00

	0.00	X
		0.00

	A	**B**	**C,E,D**	**F**
A	0.00	X	X	X
B	0.71	0.00	X	X
C,E,D	0.79	1.08	0.00	X
F	1.92	2.36	0.72	0.00

	A,B	**C,E,D**	**F**
A,B	0.00	X	X
C,E,D	0.79	0.00	X
F	1.92	0.72	0.00

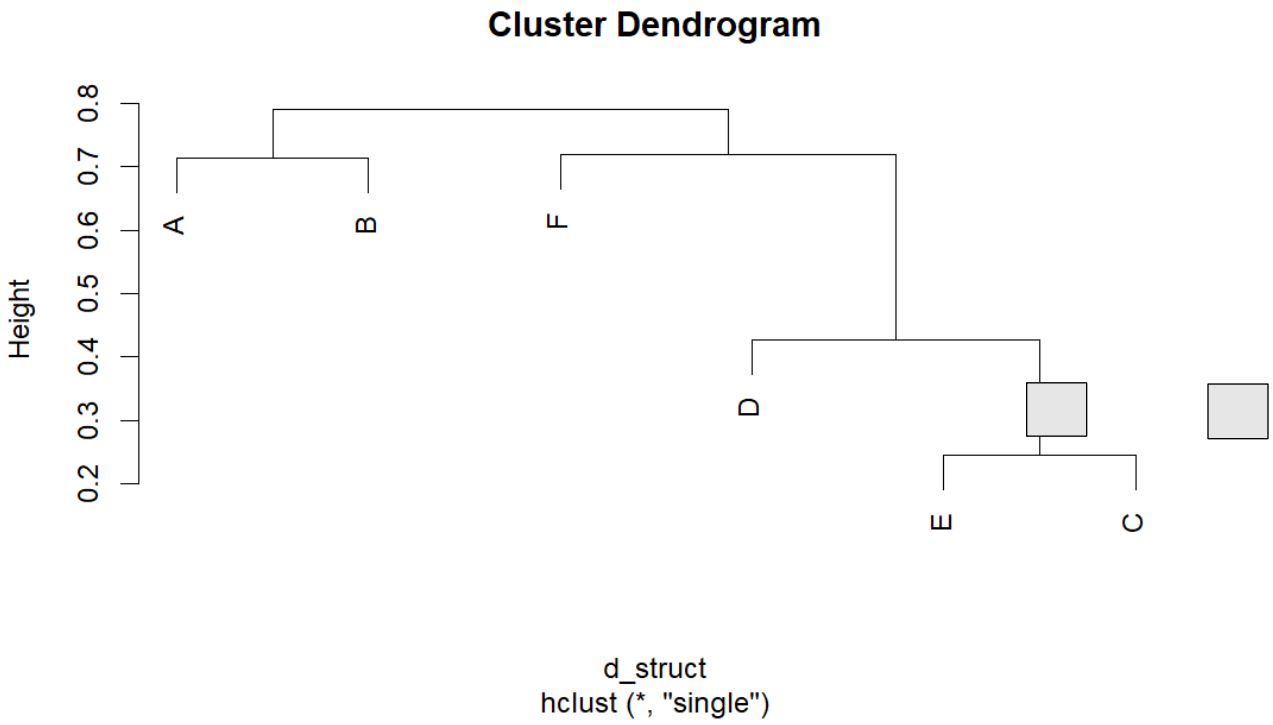
	A,B	**C,E,D,F**
A,B	0.00	X
C,E,D,F	0.79	0.00

(d) (2 points) Complete the dendrogram below by labeling the observations.

[Note: please click into each node to label it or type the node labels from left to right.]

Candidates performed well on this task overall. The first diagram below is unchanged from what was provided with the exam, and the second diagram is completed with the correct labels. Labels A and B could be assigned to the two leftmost nodes in either order.

ANSWER:



Task 6 – (12 *points*)

(a) (2 *points*) Describe the differences between using weights and offsets in an ordinary least squares model.

Candidate performance was mixed on this task. Most strong responses included examples to articulate the difference between weights and offsets, although this was not required for full credit.

ANSWER:

Offsets are used to include a variable directly in the linear predictor formula without a coefficient, or equivalently fixing the coefficient at 1. They are used when a variable has a known relationship with the target variable.

Weights are used to give certain observations greater importance in the model than other observations. For example, we can use weights to cause the model to fit better to observations with a large value for ELECTRICITY_ACCOUNTS.

(b) (3 points) Explain the differences between using a variable as a weight versus a predictor variable in the context of a generalized linear model (GLM).

Candidate performance was mixed on this task. Many candidates copied their answer from part (a). These responses only received credit when they included relevant information for this task.

ANSWER:

When a variable is used as a weight in a GLM, it adjusts the influence of each observation in the model. Observations with higher weights have more influence on the estimation of model parameters. The weight is only used in the model fitting process and is not used in the formula for calculating predictions.

When a variable is used as a predictor variable in a GLM, it is included in the prediction formula to explain variations in the target variable.

(c) (2 points) Compare and contrast ROC and AUC in the context of model performance evaluation.

Candidate performance was mixed on this task. Partial credit was awarded for descriptions of the two without comparing and contrasting. Most full credit responses identified that both metrics measure performance of a classification model and the key difference that ROC is a full curve of values whereas AUC is a single number incorporating values from the full ROC curve.

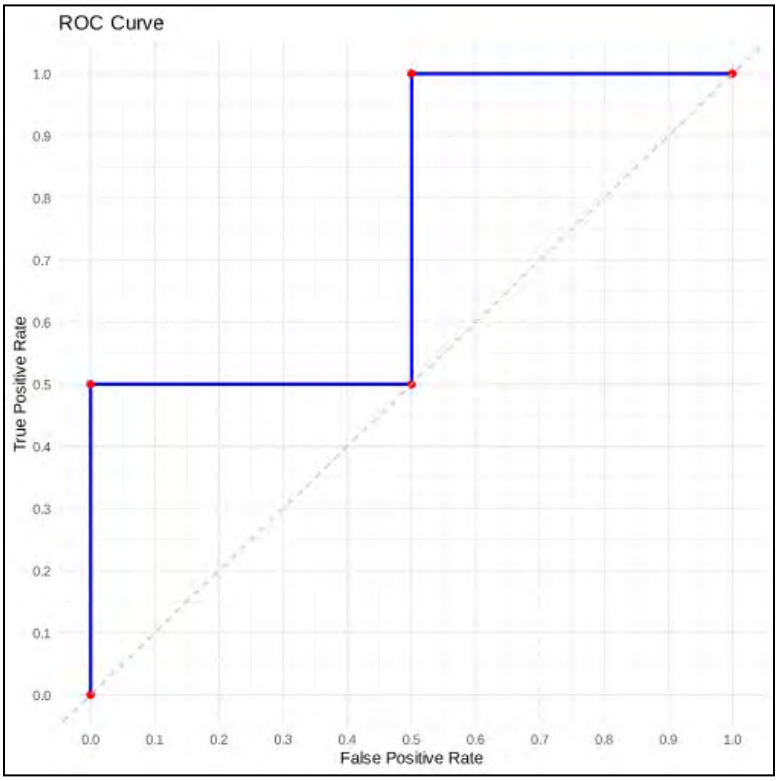
ANSWER:

Both ROC and AUC are used to evaluate the performance of binary classification models.

The ROC curve is a graphical representation of a classifier's performance across all classification thresholds. It plots the True Positive Rate (TPR, or sensitivity) against the False Positive Rate (FPR, or 1-specificity).

The AUC is a single scalar value representing the overall performance of the classifier. It is the area under the ROC curve.

You are provided with an ROC curve below:



(d) (1 points) Calculate AUC. Show your work.

Candidates performed very well on this task, with most candidate receiving full credit. A common mistake was calculating AUC as 0.25, the area between the curve and the diagonal.

ANSWER:

$$AUC = 0.5 * 0.5 + 1 * 0.5 = 0.75$$

Your client wants to understand the factors influencing per account natural gas usage (THERMS_PER_ACCOUNT). Your assistant prepares the data on total natural gas usage (TOTAL_THERMS), the number of gas accounts (GAS_ACCOUNT), and other variables such as natural gas usage in January (THERM_JAN), natural gas usage in July (THERM_JUL), and building type (BUILDING_TYPE). Your manager suggests building a logistic regression model to identify high energy users, and use ROC and AUC as evaluation metrics.

You consider using GAS_ACCOUNT as a weight variable in a generalized linear model (GLM) to better understand its impact on high natural gas usage.

Your assistant creates a binary variable HIGH_THERMS_PER_ACCOUNT to identify high natural gas accounts, and builds two GLMs, one model uses GAS_ACCOUNT as a weight, but not as a variable, and the other is unweighted and includes GAS_ACCOUNT as a variable. You are provided with model summaries and ROC curves for the 2 models.

Model 1:

```
Call:
glm(formula = HIGH_THERMS_PER_ACCOUNT ~ THERM_JAN + THERM_JUL +
    BUILDING_TYPE + AVG_STORIES + AVG_BLDG_AGE, family = binomial(link = "logit"),
    data = train_data, weights = GAS_ACCOUNT)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -7.227e+00  6.779e-02 -106.61   <2e-16 ***
THERM_JAN                4.025e-04  1.836e-06  219.24   <2e-16 ***
THERM_JUL               -1.264e-04  2.012e-06  -62.80   <2e-16 ***
BUILDING_TYPEResidential  8.325e+00  6.736e-02  123.60   <2e-16 ***
AVG_STORIES             -3.315e+00  1.319e-02 -251.27   <2e-16 ***
AVG_BLDG_AGE             1.329e-02  1.919e-04   69.28   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 508382  on 44755  degrees of freedom
Residual deviance: 319697  on 44750  degrees of freedom
  (2097 observations deleted due to missingness)
AIC: 319709

Number of Fisher Scoring iterations: 9
```

Model 2:

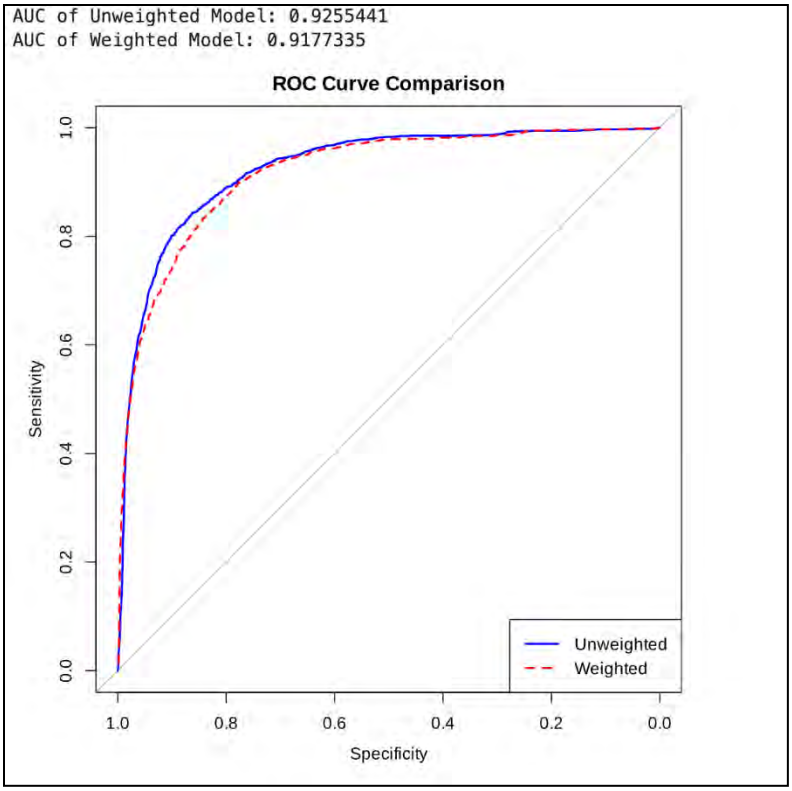
```
Call:
glm(formula = HIGH_THERMS_PER_ACCOUNT ~ THERM_JAN + THERM_JUL +
    BUILDING_TYPE + AVG_STORIES + AVG_BLDG_AGE + GAS_ACCOUNT,
    family = binomial(link = "logit"), data = train_data)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -5.577e+00  2.261e-01 -24.670  < 2e-16 ***
THERM_JAN                 1.913e-04  6.397e-06  29.904  < 2e-16 ***
THERM_JUL                 3.504e-05  5.809e-06   6.031 1.63e-09 ***
BUILDING_TYPEResidential  5.848e+00  2.249e-01  26.008  < 2e-16 ***
AVG_STORIES              -3.494e+00  6.193e-02 -56.427  < 2e-16 ***
AVG_BLDG_AGE              1.113e-02  7.710e-04  14.437  < 2e-16 ***
GAS_ACCOUNT               9.834e-02  2.177e-03  45.175  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29477  on 44755  degrees of freedom
Residual deviance: 18414  on 44749  degrees of freedom
  (2097 observations deleted due to missingness)
AIC: 18428

Number of Fisher Scoring iterations: 9
```



(e) (3 points) Interpret the model results in the context of AIC and AUC for each model.

Candidate performance was mixed on this task. Full-credit responses interpreted both AIC and AUC and related these differences to the inclusion of weights in Model 1.

ANSWER:

AUC evaluates the model's ability to distinguish between classes (e.g., true positive rate vs. false positive rate). The two curves are nearly identical, indicating that the models have equal ability to distinguish between classes.

AIC is a measure of the goodness of fit of the model, penalized for the number of parameters. It reflects the likelihood of the model given the data, with a penalty for complexity. Weights can significantly affect the log-likelihood calculation in the model. If weights amplify noise or reduce the influence of informative data points, the log-likelihood decreases, leading to a higher AIC. Hence, the lower AIC for Model 2 does not indicate it is superior.

(f) (1 points) Recommend a model to your client. Justify your answer.

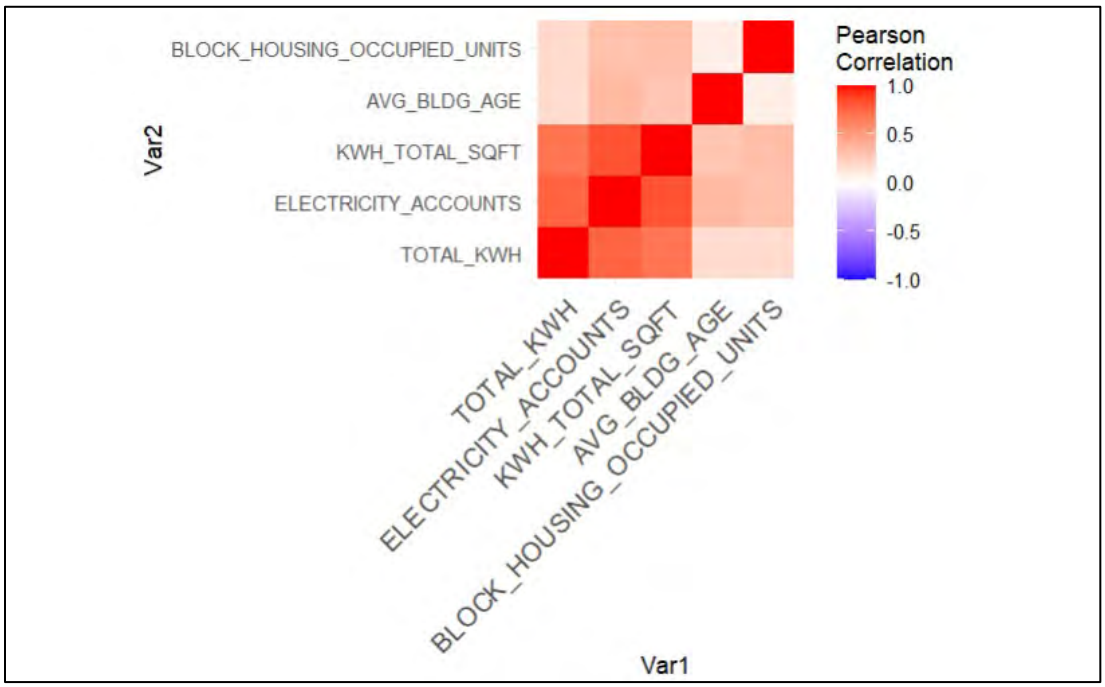
Candidates performed well on this task. A sound recommendation based on model performance was sufficient for full credit, although some candidates also discussed the additional complexity of incorporating a weighting variable.

ANSWER:

Model 2. From model output, adding GAS_ACCOUNT as a weight doesn't improve the overall fit, nor improve the AUC to better identify high gas users.

Task 7 – (3 *points*)

Your assistant is modeling the TOTAL_KWH of a block as the target variable. They want to examine the relationships between the variables being considered and has built a correlation heat map below.



(a) (*2 points*) Interpret the graphic and identify which independent variables exhibit collinearity.

Candidates performed well on this task. The most common mistake was interpreting TOTAL_KWH as a predictor variable rather than the target variable.

ANSWER:

The chart displays a correlation heat map visualizing the relationship between pairs of variables. From the chart we can assess which predictor variables have potential explanatory power for the TOTAL_KWH, the target variable. We can also identify potential multicollinearity concerns with the predictor variables.

The variables that are most correlated with the target variable, TOTAL_KWH, are ELECTRICITY_ACCOUNTS and KWH_TOTAL_SQFT. These variables look like they will be very predictive if included in a model. However, ELECTRICITY_ACCOUNTS and KWH_TOTAL_SQFT are also highly correlated with each other, indicating that they may cause collinearity problems if used together in a model.

(b) (*1 point*) Recommend and justify one enhancement to improve the chart.

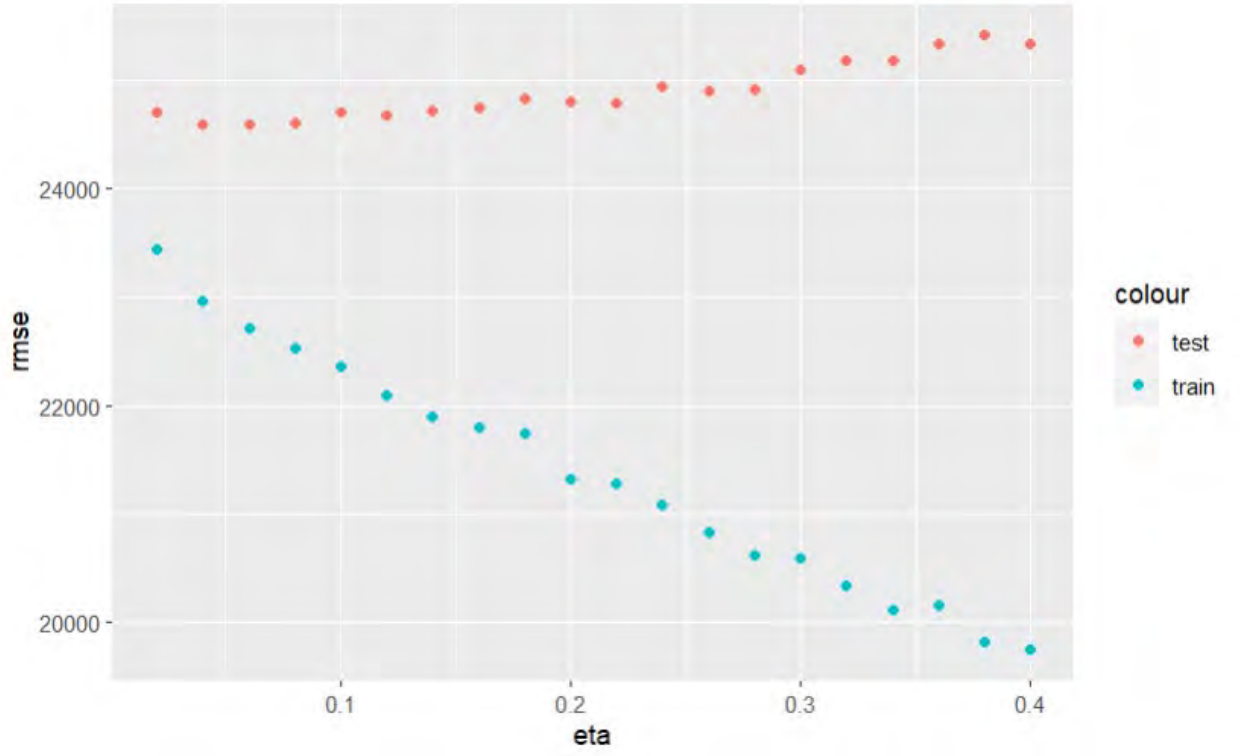
Candidates performed well on this task. Most full-credit answers identified the shortcoming of using color hue/saturation to encode data and made a recommendation that improves the readability of the values.

ANSWER:

The correlations values in the chart are encoded using color gradient. I recommend adding a label in each box in the chart with the value of the correlation since it is hard to accurately read small differences in values from gradients.

Task 8 – (4 *points*)

Your assistant decides to build a boosted tree and is tuning the model hyperparameter for an optimal learning rate e. Your assistant has divided the data into a training set with 80% of the data and a testing set with 20% of the data. They notice the model evaluation metric (RMSE) on the train set (without cross validation) and test set behave differently when changing the learning rate.



(a) (*2 points*) Explain why increasing the learning rate would result in such a difference in performance on the train vs. test set for a boosted tree.

Candidate performed well on this task. Full-credit responses discussed how overfitting impacts model performance on unseen data differently from the data used to train the model.

ANSWER:

The learning rate corresponds to how quickly the model adapts to the training data. Boosting builds sequential trees where it tries to minimize the loss value from the prior tree, which can result in lower amount of bias in the model development process. By applying a higher learning rate, we would expect the model to overfit, which would also cause the RMSE on the training data to perform better.

Conversely, we can observe the impact of the overfitting by observing the RMSE on the test set, which indicates an opposite trend. That is, as we increase the rate of learning, which reduces model bias, we end up with worse results and increased variance.

Your assistant would like to set a value for the learning rate based on the results from the test set in the chart above.

(b) (*2 points*) Critique the assistant's proposed method of hyperparameter tuning. Recommend and justify an alternative approach.

Candidate performance was mixed on this task. Full-credit responses articulated the issue of data leaking from the test set.

ANSWER:

We are tuning the hyperparameters based on the test data, which is not advisable. Although the results of this tuning experiment may guide our choice of an optimal hyper parameter, we are leaking information from the test set into the model building process. A better approach is to use k-fold cross validation on the training data to tune the hyperparameters. This allows us to keep the test data set aside so that we can use it to assess the final model's performance on truly unseen data.

Task 9 – (8 *points*)

Your manager is working on a project for the city of Chicago to measure the impact of weather and climate on energy use. Your manager asks your assistant to prepare a graphic overview of monthly weather patterns.

Your assistant isn't sure how best to aggregate the weather variables, which are captured on a daily basis, into monthly variables. The description of each of the weather variables is copied below from the data dictionary.

- TMAX_FAHRENHEIT: Max temperature recorded during the day.
- TMIN_FAHRENHEIT: Min temperature recorded during the day.
- PRECIPITATION_INCHES: Inches of precipitation during the day.
- SNOW_FALL_INCHES: New snowfall during the day.
- SNOW_DEPTH_INCHES: Snow depth reported at 7 am each day.

Your manager wants to be able to use these monthly weather variables in linear models and for the interpretation of the variables to be intuitive.

(a) (*2 points*) Recommend whether each of the following weather variables should be aggregated based on taking the average of the daily values or the sum of the daily values; briefly justify your recommendation.

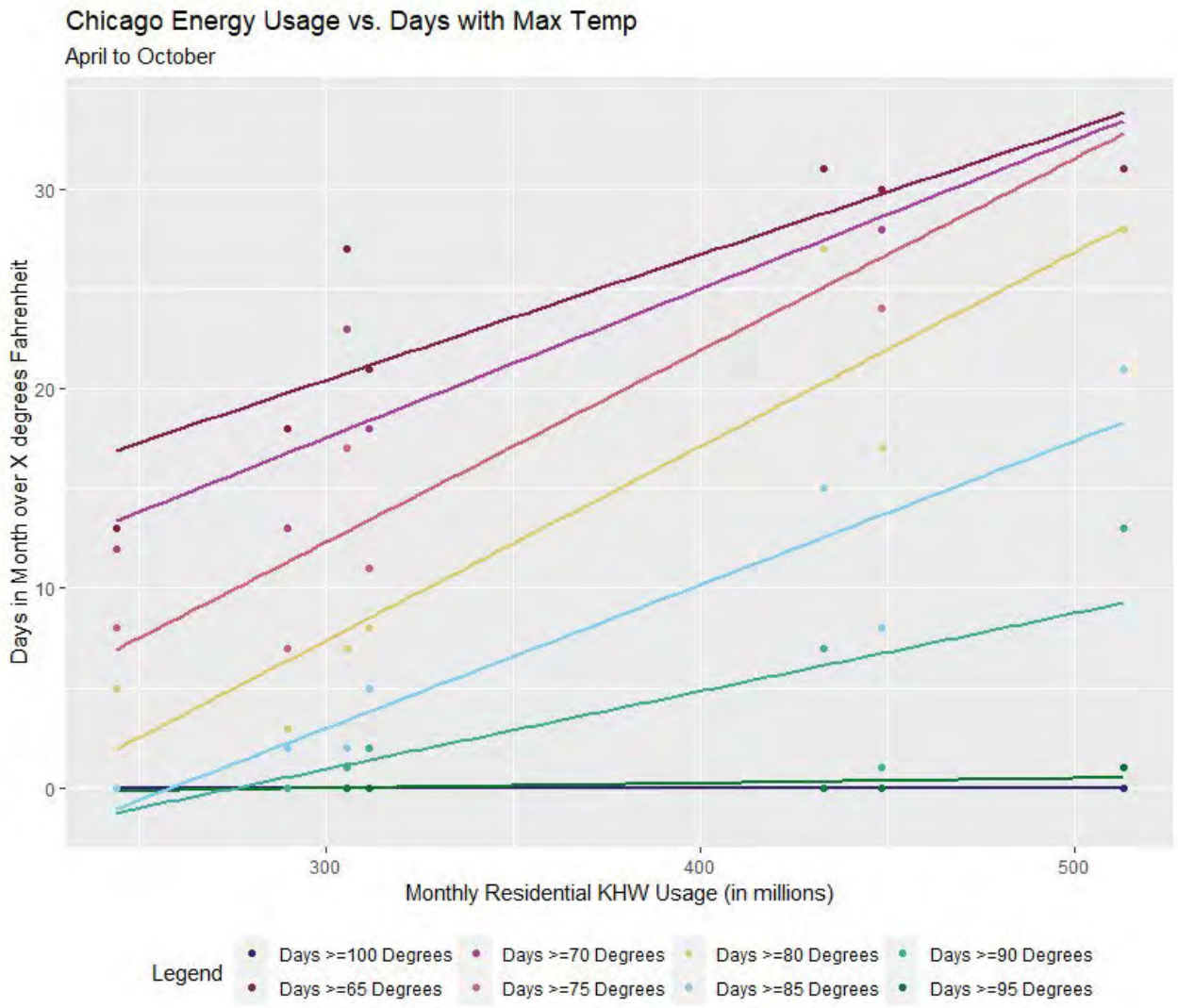
Candidates struggled with this task. A clear recommendation and sound justification were both required for full credit. Most full-credit responses recommended averaging for both variables.

ANSWER:

TMAX_FAHRENHEIT
Recommended Monthly Aggregation: Recommend aggregating based on the average of the daily values.
Justification: Using average temperature will be more intuitive than using the sum of daily maximum temperatures. While using this value in a linear model will lead to the same outcome, after scaling the coefficient, the marginal impact of average monthly temperature being higher or lower by one degree will be easier to interpret in the model results. You also avoid the issue of months with more days having a higher sum of monthly temperatures due to the length of the month.

SNOW_DEPTH_INCHES
Recommended Monthly Aggregation: Recommend aggregating based on the average of the daily values.
Justification: Snow depth inches is an accumulative measure. Measuring the snow depth each day and summing those values may make it difficult to interpret the meaning of the variable since additional snow depth on one day could affect snow depth on future days.

Your assistant observes that residential energy usage in terms of kilowatt hours peaks in April-October and hypothesizes that this is due to air conditioning usage and higher temperatures. Your manager agrees, but isn't sure whether the relationship is due to the average temperature during the month or the number of days in a month that the temperature exceeds a specific level. Your assistant graphs the number of days that the maximum temperature exceeds different levels in each month against monthly residential Kilowatt hours for April through October.



(b) (*2 points*) Briefly summarize the relationship between monthly residential KWH usage and days in a month exceeding a specific temperature based on the graph above.

Candidate performance was mixed on this task. Full-credit responses noted the positive correlation at most thresholds, and also observed that this pattern does not hold for days >= 95 or 100 degrees.

ANSWER:

There is a positive relationship between the number of days in a month exceeding a specific temperature and monthly residential KWH usage in millions for temperatures between 70 degrees and 90 degrees Fahrenheit. Due to a lack of months with days over 95 degrees and no months having days over 100 degrees, it isn't possible to tell the relationship between the number of days in a month exceeding those temperatures and monthly residential energy use.

Your assistant also provides the table of values for the chart above.

Month	Days over 100 degrees	Days over 95 degrees	Days over 90 degrees	Days over 85 degrees	Days over 80 degress	Days over 75 degrees	Days over 70 degrees	Days over 65 degrees	Monthly Residential KWH Usage (in millions)
April	0	0	0	0	5	8	12	13	244
May	0	0	2	5	8	11	18	21	312
June	0	0	1	8	17	24	28	30	449
July	0	1	13	21	28	31	31	31	513
August	0	0	7	15	27	31	31	31	433
September	0	0	1	2	7	17	23	27	306
October	0	0	0	2	3	7	13	18	290

(c) (*2 points*) Recommend and justify one improvement in your assistant's analytical approach and one improvement in your assistant's graph design. Your answer should be based on the graph and table provided.

Candidate performance was mixed on this task. Full credit was awarded for recommending improvements to both the analysis approach and the graph design, where both recommendations related to the plot and table provided.

ANSWER:

I would recommend excluding the Days over 100 degrees and Days over 95 degrees variables since there is very little variation across months (no variation for Days over 100 degrees. Also, consider combining some of the degree ranges, for example doing every 10 degrees instead of every 5 degrees.

I would recommend adding labels to the graph to indicate that each vertical set of points represents a single month's energy utilization and number of days over a specific number of degrees. I would also remove the lines for Days over 95 degrees and Days over 100 degrees since they don't aid in interpreting the relationship between the variables. Also consider labeling the slopes of the lines.

Your manager is interested in a deeper understanding of the impact of weather and climate on energy usage in Chicago. You are concerned that the current data isn't sufficient to support some of your manager's questions. For each question below:

(d) (*2 points*) Explain whether or not the analysis can be supported by the current data. If the data is available, state which variables you would use. If the data is not sufficient, state the additional data you would need to collect.

Candidate performance was mixed on this task. Full-credit answers included a clear explanation of whether each analysis can be supported, including variables to be used and additional data if not supportable.

ANSWER:

Which decade of residential building age is most efficient in terms of energy use during June, July, and August?

Do you have sufficient data to perform the analysis? Yes.

If so what variables would you use? If not, what additional data would you need to request?
I would use the variables TMAX_FAHRENHEIT and TMIN_FAHRENHEIT to adjust for temperature. I would use BUILDING_TYPE to filter to just residential buildings. I would use KWH_JUN, KWH_JUL, KWH_AUG to focus on energy use in those months and I would use BUILDING_AGE to reflect the age of the building.

Do communities in Chicago nearer to Lake Michigan experience lower average temperatures and correspondingly require less energy during June, July, and August?

Do you have sufficient data to perform the analysis? No.

If so what variables would you use? If not, what additional data would you need to request?
We don't currently have temperature data by neighborhood, just a single daily record for all of Chicago. I would need neighborhood-level temperature data to answer this question. We would also need the distance to Lake Michigan for each community.

Task 10 – (11 *points*)

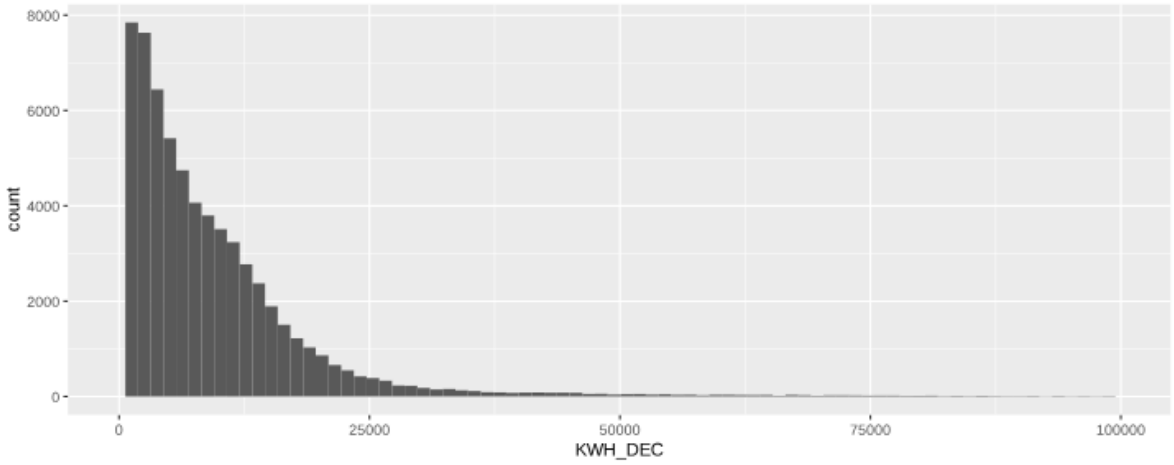
(a) (2 point) Describe the key assumptions of the generalized linear model (GLM).

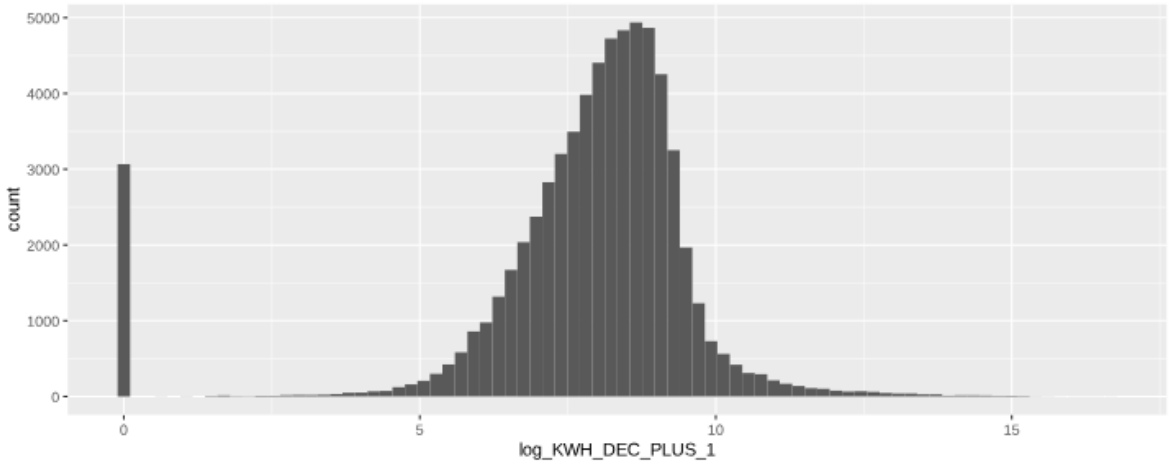
Candidates struggled with this task, with many candidates unable to articulate any GLM assumption.

ANSWER:

- All observations in the data are independent.
- The distribution of the target variable is a member of the linear exponential family.
- The model prediction is a transformation of a linear combination of the predictor variables; $\mu = g^{-1}(\eta)$, $\eta = X\beta$, where g is called the link function, and g^{-1} is its inverse.

Your client wants to use the previous months' electricity usage to predict December's utilization by building type. Your manager suggests making transformations of the KWH variables. Your assistant took December utilization (KWH_DEC) and made a histogram without transformation, and with log(KWH_DEC+1) transformation. Your assistant suggests to use the log(KWH_DEC+1) transformation.





(b) (3 points) Explain the pros and cons of your assistant's suggested variable transformation.

Candidates struggled with this task. Full-credit responses articulated at least one pro and con.

ANSWER:

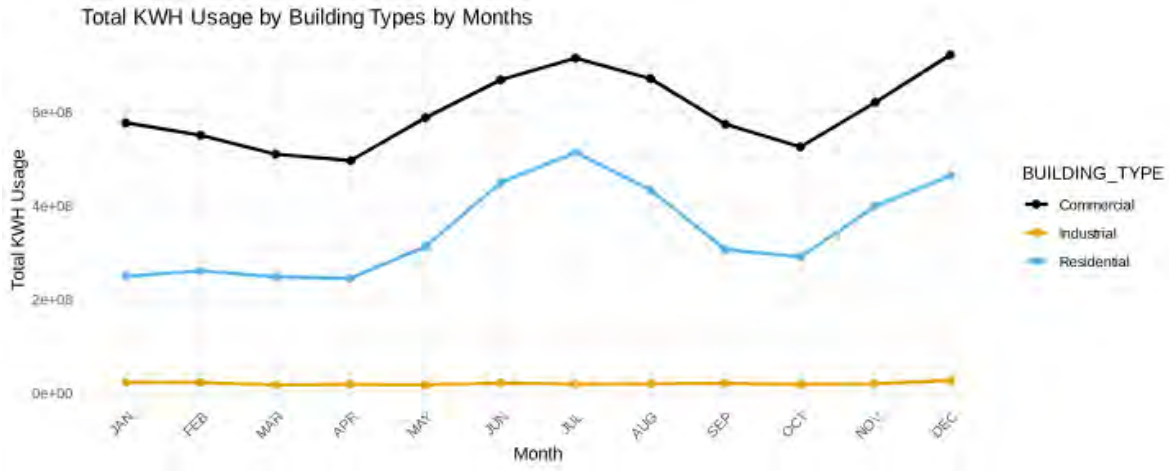
Pros:

- The transformation removes right-skew from the variable, making its distribution more symmetric.
- The transformation reduces the impact of outliers.

Cons:

- The transformation makes it more difficult to interpret model coefficients.
- Applying a transformation will not necessarily improve model performance.
- The transformation has a spike at zero.

Your assistant provides you with the total KWH usage by building type by month plot below and points out the cyclic pattern for commercial and residential building types. Your assistant suggests applying seasonality to the entire dataset.



(c) (2 points) Justify your assistant's suggestion and recommend a way to model this cyclic pattern.

Candidate performance was mixed on this task. Recognizing that the seasonality effect should be modeled differently for Industrial than the other building types was required for full credit. Credit was awarded for sound recommendations of alternative methods to model seasonality, including trigonometric transformations.

ANSWER:

I agree with my assistant on applying seasonality to the data since both commercial and residential building types are showing consistent rise and fall in energy usage at the same times in the year, indicating the data has a seasonal component. However, industrial usage does not show a clear cyclic pattern.

I recommend Including categorical features for months to account for monthly variations in energy usage, including an interaction term with the Industrial building type.

Your assistant has built two models to predict December electricity usage using months and BUILDING_TYPE variables.

Model 1: Each row corresponds to an observation for a specific year. The columns include variables for different months (e.g., KWH_JAN, KWH_FEB, ..., KWH_NOV).

Model 2: Each row corresponds to an observation at a specific time point (e.g., month). Separate columns are used for the Month variable and the electricity usage (e.g., KWH).

You are provided with the model summary output. Your assistant also pointed out that BUILDING_TYPE is statistically significant in Model 2 but not Model 1.

Model 1:

```
Call:
glm(formula = KWH_DEC ~ BUILDING_TYPE + KWH_JAN + KWH_FEB + KWH_MAR +
    KWH_APR + KWH_MAY + KWH_JUN + KWH_JUL + KWH_AUG + KWH_SEP +
    KWH_OCT + KWH_NOV, family = gaussian(link = "identity"),
    data = energy_data)

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.626e+02  1.241e+02   1.310    0.190
BUILDING_TYPEIndustrial    4.190e+03  3.153e+03   1.329    0.184
BUILDING_TYPEResidential -7.015e+00  1.425e+02  -0.049    0.961
KWH_JAN                    1.693e-01  4.900e-03  34.549   <2e-16 ***
KWH_FEB                    6.850e-01  7.175e-03  95.475   <2e-16 ***
KWH_MAR                   -1.959e-01  7.631e-03 -25.670   <2e-16 ***
KWH_APR                   -2.185e-01  8.005e-03 -27.292   <2e-16 ***
KWH_MAY                   -5.189e-01  7.480e-03 -69.364   <2e-16 ***
KWH_JUN                    7.977e-02  5.111e-03  15.607   <2e-16 ***
KWH_JUL                    2.228e-01  5.230e-03  42.611   <2e-16 ***
KWH_AUG                   -1.049e-01  6.616e-03 -15.855   <2e-16 ***
KWH_SEP                    1.778e-01  6.030e-03  29.486   <2e-16 ***
KWH_OCT                   -2.068e-01  6.641e-03 -31.142   <2e-16 ***
KWH_NOV                    1.003e+00  4.706e-03 213.086   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 250210309)

    Null deviance: 2.6726e+15  on 66102  degrees of freedom
Residual deviance: 1.6536e+13  on 66089  degrees of freedom
  (871 observations deleted due to missingness)
AIC: 1465896

Number of Fisher Scoring iterations: 2
```

Model 2:

```
Call:
glm(formula = KWH ~ BUILDING_TYPE + Month, family = gaussian(link = "identity"),
    data = energy_data_long)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                34312.3      708.5  48.430  < 2e-16 ***
BUILDING_TYPEIndustrial   717419.8     9238.8  77.653  < 2e-16 ***
BUILDING_TYPEResidential -29122.4      422.0 -69.017  < 2e-16 ***
MonthKWH_FEB                -228.2      896.9  -0.254  0.79916
MonthKWH_MAR               -1103.4      896.9  -1.230  0.21864
MonthKWH_APR               -1346.8      896.9  -1.502  0.13320
MonthKWH_MAY                1042.5      896.9   1.162  0.24513
MonthKWH_JUN                4402.6      896.9   4.908 9.18e-07 ***
MonthKWH_JUL                6035.2      896.9   6.729 1.71e-11 ***
MonthKWH_AUG                4179.3      896.9   4.660 3.17e-06 ***
MonthKWH_SEP                 785.0      896.9   0.875  0.38149
MonthKWH_OCT                -215.3      896.9  -0.240  0.81032
MonthKWH_NOV                2895.3      896.9   3.228  0.00125 **
MonthKWH_DEC                5504.4      896.9   6.137 8.41e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 26589415620)

    Null deviance: 2.1393e+16  on 793235  degrees of freedom
Residual deviance: 2.1091e+16  on 793222  degrees of freedom
  (10452 observations deleted due to missingness)
AIC: 21291784

Number of Fisher Scoring iterations: 2
```

(d) (2 points) Explain the reason that BUILDING_TYPE is statistically significant in Model 2 but not Model 1 from the model summary outputs.

Candidates struggled with this task. Many responses were incomplete, possibly resulting from time constraints. Full-credit responses provided an explanation for the different t-values between the models.

ANSWER:

In Model 1, KWH_NOV has a large t-value, meaning KWH_DEC is highly correlated with KWH_NOV. Since KWH_NOV captures large data variation, and potential collinearity with other months, it makes BUILDING_TYPE variable less significant.

In Model 2, instead of predicting December KWH, the model use month and KWH as two separate variables. It therefore breaks the correlation among the months and makes BUILDING_TYPE statistically significant.

(e) (2 point) Recommend a model to your manager and justify your answer.

Candidate performance was mixed on this task. Full-credit responses included a sound recommendation and demonstrated an understanding the difference in how the two models are constructed.

ANSWER:

I recommend Model 1. Model 1 incorporates information on energy usage from each earlier month in the year into the prediction. This allows the model to capture relationships between December and prior months in the calendar year.

Model 2 does not consider prior months in its predictions. It takes a simpler approach, applying a single Month coefficient in each prediction.

We see that Model 1 has much smaller AIC and residual deviance, indicating better overall model performance.