Exam PA October 12, 2022 Project Statement

This model solution is provided so that candidates may better prepare for future sittings of Exam PA. It includes both a sample solution, in plain text, and commentary from those grading the exam, in italics. In many cases there is a range of fully satisfactory approaches. This solution presents one such approach, with commentary on some alternatives, but there are valid alternatives not discussed here.

General Information for Candidates

This examination has 12 tasks numbered 1 through 12 with a total of 100 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 (and related data files) and data dictionary 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. An .Rmd file accompanies this exam and provides useful R code for importing the data and, for some tasks, additional analysis and modeling. There are five datasets used in this exam. They are all subsets of a larger dataset that is not given to candidates. The .Rmd file has a chunk for each task. Each chunk starts by reading in one or more data files into one or more dataframes that will be used in the task. This ensures a common starting point for candidates for each task and allows them to be answered in any order. When the datafile is read, the variables it contains are assigned a type (e.g., "numerical," "factor"). The code that assigns variable types is easily changed (e.g., if month is read in as "numeric" but you want to treat it as a factor).

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. Where code, tables, or graphs from your own work in R is required, it should be copied and pasted into this Word document.

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. When "for a general audience" is specified, write for an audience **not** familiar with analytics acronyms (e.g., RMSE, GLM, etc.) or analytics concepts (e.g., log link, binarization).

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.

It is not required to upload your .Rmd file or other files used in determining your responses, as needed items from work in R will be copied over to the Word file as specified in the subtasks.

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

Business Problem

Your boss recently started a consulting firm, PA Consultants, specializing in predictive analytics. You and your assistant are the only other employees. Your boss informs you that a local politician from Baton Rouge, Louisiana, USA has hired your firm.

Baton Rouge, a city of about 230,000 residents, is the capital of the state of Louisiana, USA.

The client is about to launch a campaign with the mottos, "Clean up Baton Rouge" and "Treat all Neighborhoods Equally – including yours!" The client wants to improve garbage and waste collection. In particular, the client cares about shortening resolution times and ensuring equitable resolution times throughout the city.

The client wants your ideas and inputs on the following.

- *Understanding time trends*
- *Seeing whether different responding departments have different resolution times for similar tasks*
- *Predicting resolution times for any type(s) of complaint.*

Your boss directs you to use a dataset[1] of public data that includes all the service requests from January, 2016 – March, 2022. There are over 300,000 service requests in this time period. Your assistant has prepared five subsets of the public data and has provided the following data dictionary that contains all the variables appearing in the subsets. Note that all variables do not appear in every subset datafile.

[1] *Source: City of Baton Rouge Parish of East Baton Rouge*.

Data Dictionary

Variable Name	Variable Values
Time.to.resolution	Days from service request to resolution
quarter	“Q1”, “Q2”, “Q3”, “Q4”; quarter of service request
month	1 to 12, month of service request
year	2016 to 2022, year of service request
year.mo	201601 to 202203, 100*year + month
weekday	“Sunday”, … ”Saturday”; day of the week for the service request
TYPEid	An id representing a specific type of service request
SERVICE.REQUEST.ID	Unique code assigned to service request
DEPARTMENT	“GROUNDS”,”BLIGHT”,”SANITATION”
LATITUDE	Latitude of service location, 30.2 to 30.6
LONGITUDE	Longitude of service location, -91.3 to -90.9
area	“N”,”W”,”D”,”LSU”; neighborhood of service location
Latitude_binned	Latitude range for binned data (geo.grid.csv only)
Longitude_binned	Longitude range for binned data (geo.grid.csv only)
Ave.time.to.resolution	Average Time.to.resolution for binned data (geo.grid.csv only)
call_count	Number of service requests for binned data (geo.grid.csv only)

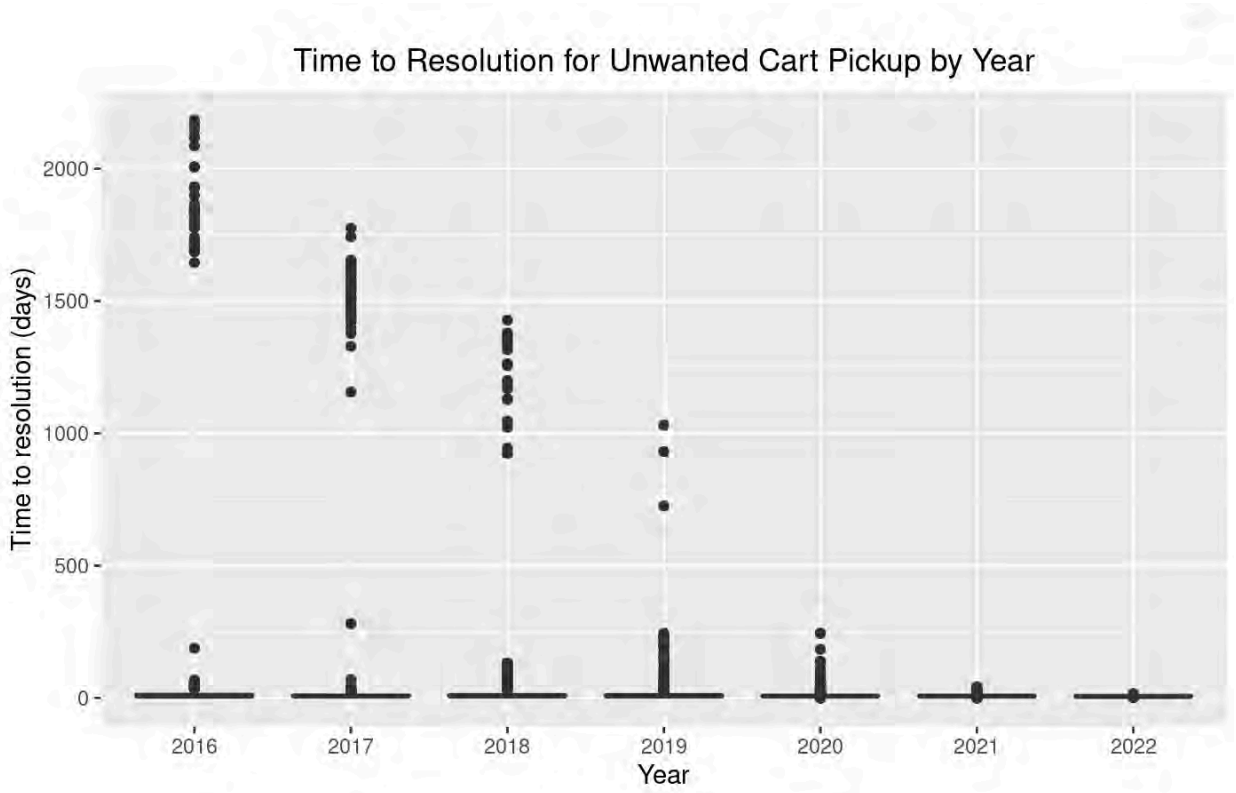
Comments

Requests for service do not appear in the dataset until they are resolved.

Task 1 (*7 points*)

Your boss asks you to review the quality of the data below. The data shows Time to Resolution for calls to pick up unwanted garbage carts. (This data is not found in any of the supplied files.)

(a) (*2 points*) Review the box plot below that your assistant made and describe an issue with the data.



Candidates received full credit for identifying outliers with very high time to resolution as an issue and describing how the outliers may arise, patterns in the outliers, or how the outliers could cause problems in addressing the business problem. A common mistake was misidentifying the outliers as the body of the distribution and stating the actual boxplot represents unreasonable zero values, when in fact, this is an artifact of the scale of the y-axis caused by the high outliers.

ANSWER:

The plot shows many outlier resolution times greater than one year. These resolution times are unreasonable for trash services. This suggests either that services were never performed or that the cases were not closed at the time service was completed.

(b) (*1 point*) List three options for handling the data issue.

Candidates received full credit for listing three distinct options that addressed the data issue. The most common mistakes were listing options to improve the graph rather than handle the data issue (e.g., using a log scale) and giving vague response (e.g. listing "further investigation" as an option).

ANSWER:

1. Remove outliers with very high time to resolution from the dataset
2. Leave the outliers in the dataset without any modification
3. Truncate the time to resolution variable

(c) (*2 points*) Select and explain which option from part (b) you would recommend.

Candidates performed well on this task overall. The most common recommendation was removing the outliers, but full credit was granted for any recommendation with a reasonable explanation.

ANSWER:

I recommend removing the data with excessive resolution times. It seems likely that the requests were not closed when the service was performed because these response times stretch over multiple years.

(d) (2 *points*) Your assistant produces the following output from a GLM. (Note your assistant redefined year as years since 2016.)

This is a relatively straightforward calculation task, and candidates performed well overall. Varying amounts of partial credit were awarded to candidates with incorrect answers. Calculation errors (e.g., missing a coefficient in the formula) were awarded more partial credit than incorrect formulas (e.g., ignoring or misapplying the link function, incorrect residual calculation).

```
[1] "Formula:"
Time.to.resolution ~ year + as.factor(month) + as.factor(TYPEid) +
    area

Call:
glm(formula = formula1, family = Gamma(link = "log"), data = df2.sanitation)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4555  -0.4824  -0.2193   0.1572   2.9248

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)               2.666173   0.007380  361.272  < 2e-16 ***
year                     -0.124969   0.001037 -120.547  < 2e-16 ***
as.factor(month)2        -0.123720   0.009119  -13.567  < 2e-16 ***
as.factor(month)3        -0.077945   0.008557   -9.109  < 2e-16 ***
as.factor(month)4         0.035228   0.008471    4.159 3.20e-05 ***
as.factor(month)5         0.093898   0.008134   11.544  < 2e-16 ***
as.factor(month)6         0.014100   0.008154    1.729   0.0838 .
as.factor(month)7         0.054114   0.008021    6.747 1.52e-11 ***
as.factor(month)8         0.020327   0.008080    2.516   0.0119 *
as.factor(month)9        -0.085676   0.008259  -10.373  < 2e-16 ***
as.factor(month)10       -0.077113   0.008562   -9.006  < 2e-16 ***
as.factor(month)11       -0.083417   0.008953   -9.317  < 2e-16 ***
as.factor(month)12       -0.136517   0.008646  -15.789  < 2e-16 ***
as.factor(TYPEid)173023  -0.637010   0.004865 -130.934  < 2e-16 ***
as.factor(TYPEid)173024  -0.233447   0.006019  -38.784  < 2e-16 ***
as.factor(TYPEid)173027  -0.274727   0.005549  -49.511  < 2e-16 ***
as.factor(TYPEid)173028  -0.144072   0.005467  -26.351  < 2e-16 ***
as.factor(TYPEid)427105  -0.830102   0.005525 -150.237  < 2e-16 ***
areaLSU                  -0.011815   0.004934   -2.395   0.0166 *
areaN                    -0.056956   0.004919  -11.579  < 2e-16 ***
areaW                    -0.022671   0.004017   -5.643 1.67e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.4794437)

    Null deviance: 87291  on 182086  degrees of freedom
Residual deviance: 62690  on 182066  degrees of freedom
AIC: 1048661

Number of Fisher Scoring iterations: 7
```

Calculate the residual for the predicted time to resolution using the values in the following table for a single observation. Show both the formula(s) used (with values substituted for variables) and the final value to two decimal places.

TYPEid	month	year	Area	Time to Resolution
173023	2	4	N	5

ANSWER:

$\hat{y} = \exp(2.66173 - 0.637010 - 0.124969 \times 4 - 0.123720 - 0.056956) = 3.85 \text{ days}$

$r = 5 - 3.85 = 1.15 \text{ days}$

Task 2 *(11 points)*

The client is interested in improving the debris collection performance.

Candidates performed well on this task overall. For full credit, both a correct, legible table, and the R code used to generate the table were required.

(a) (2 points) Create a table showing number of observations by year and month. Paste the R code and the table below.

ANSWER:

R Code:

```
table(df.debris$year, df.debris$month)
```

Table:

```
 1 2 3 4 5 6 7 8 9 10 11 12
2016 239 330 457 483 633 883 603 372 707 910 706 521
2017 716 665 1012 912 907 763 854 851 968 832 672 427
2018 385 613 2030 1806 1527 1455 1436 1275 1025 1042 595 606
2019 729 754 1113 1547 1737 1857 1836 1609 1172 722 542 602
2020 566 484 895 881 1366 1367 1601 1125 1024 763 554 695
2021 361 225 918 607 1112 2321 1248 1889 960 1032 693 640
2022 404 241 100 0 0 0 0 0 0 0 0 0
```

(b) (2 *points*) Recommend which time period you will choose to use for your analysis (in terms of years and months). Justify your recommendation.

Candidates performed well on this task overall. For full credit, candidates needed to address incomplete 2022 data and provide a justification grounded in the business problem.

ANSWER:

I recommend using all months of data from years 2018-2021. Starting with 2018 gives four years of data, which is enough to see recent trends. I recommend not using 2022 data since there is not data for all months, and we are only excluding 745 observations by removing 2022 data.

Your boss told your assistant to use stratified sampling when separating the chosen dataset into a training dataset and a testing dataset.

(c) (2 *points*) Discuss the benefits of stratified sampling.

Candidate performance was mixed on this task. Partial credit was awarded for a definition of stratified sampling, but a clear discussion of the benefits was required for full credit.

ANSWER:

Stratified sampling results in test and train datasets that are similar with respect to the stratification variables. To the extent that the stratification variables are related to the target variable, stratification will allow for more precise train and test estimates. Not stratifying on important predictor variables would add variance to the model because it would be fit to the segmentation of the training data, which is similar to overfitting to noise in the dataset. The test dataset would have a different segmentation, and therefore the model may not fit the test data as well as the train data.

Your assistant has stratified the entire dataset, based on month and year, and divided it into train and test datasets. You need to remove any observations that you decided not to use in (b).

(d) (*2 points*) Remove the observations that you decided in (b) not to use from the train and test datasets. Copy the code to adjust datasets.

This is a straightforward question on R coding. The majority of candidates received full credit.

ANSWER:

Code to adjust datasets:

```
debris_train <-debris_train[debris_train$year < 2022,]
debris_train <-debris_train[debris_train$year > 2017,]
debris_test <-debris_test[debris_test$year < 2022,]
debris_test <-debris_test[debris_test$year > 2017,]
```

Your assistant has prepared glm1 and glm2. Run the .Rmd file to fit the models.

(e) (3 *points*) State the better of the two models, based on RMSE. Copy the code (i.e., the glm command, and any further lines of code) for both of the models that you used to make the choice.

Candidate performance was mixed on this task. Many candidates received some amount of partial credit due to coding errors, incorrectly stated that higher RMSE indicates better model performance, or identifying a better model based on criteria other than RMSE.

ANSWER:

Model choice (erase one): Poisson GLM

RMSE for Gamma GLM: 60.2934

RMSE for Poisson GLM: 55.3739

Code to calculate Gamma GLM RMSE:

```
glm1.time <- predict(model.glm1, newdata=debris_test, type =
"response")
```

```
RMSE.1 <- sqrt(mean((debris_test$Time.to.resolution - glm1.time)^2))
```

Code to calculate Poisson GLM RMSE:

```
glm2.time <- predict(model.glm2, newdata=debris_test, type =
"response")

RMSE.2 <- sqrt(mean((debris_test$Time.to.resolution - glm2.time)^2))
```

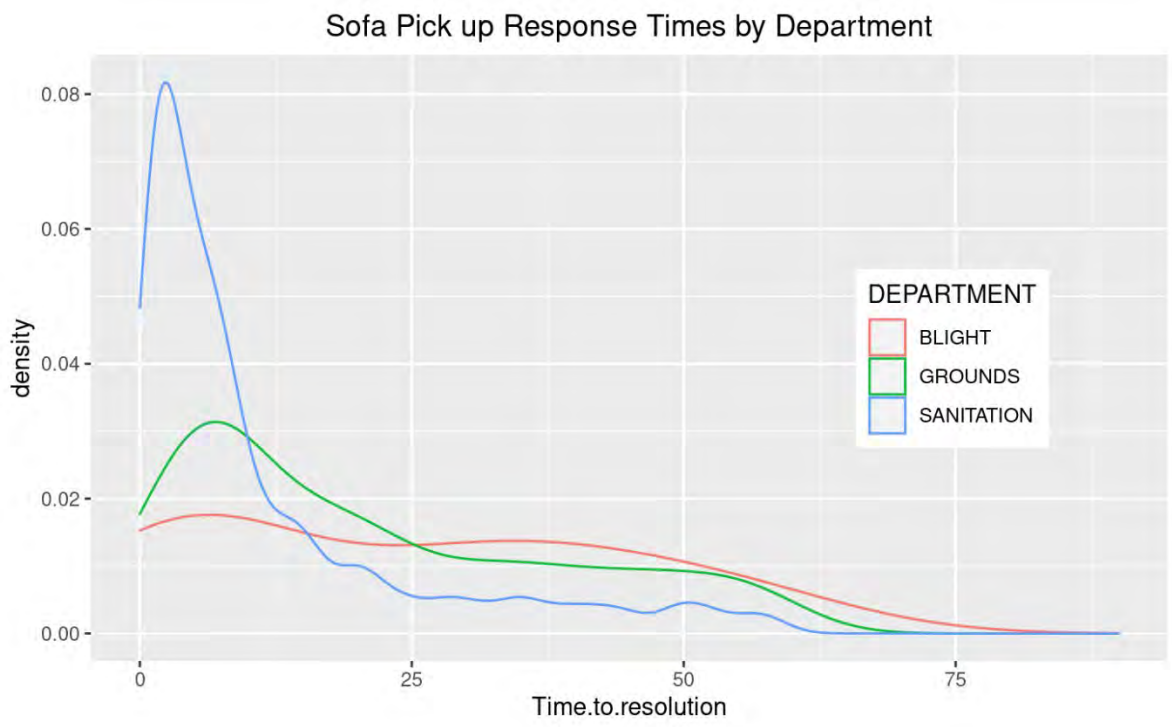
Task 3 (*12 points*)

Your boss wants to understand how the distribution of Time.to.resolution for requests to pick up sofas differs between departments. There are three departments that pick up sofas: Blight, Grounds, and Sanitation.

(a) (*3 points*) Create a single plot of Time.to.resolution by department for the provided sofa data.

Candidates performed well on this task overall, with the majority of candidates receiving full credit. Full credit was awarded for plots that were properly labeled, informative, and presented the correct data.

ANSWER:



(b) (*2 points*) Describe the observed differences by department.

Candidates performed well on this task overall. Full credit was awarded for responses that identified at least two differences across departments, with most full credit responses commenting on how the center and skewness of Time.to.resolution varies across departments.

ANSWER:

The sanitation department has the lowest response times, with a much higher density of response times below 10 than the other departments. Sanitation also has the least right skewness of the three departments.

By contrast, blight has nearly uniform density for response times up to 50, with densities decreasing slowly at response times above 50. Blight has the most skewed distribution, with highest average response times.

The grounds department is in between the other two departments in terms of overall distribution.

Your boss asks you to fit a GLM to the sofa data.

(c) (2 *points*) Explain why a Gamma distribution is a more appropriate choice of distribution than Gaussian to model Time.to.resolution.

Candidate performance was mixed on this task. Full credit responses connected the characteristics of the data to the distributions and included some discussion of the implications to fitting a GLM.

ANSWER:

Time.to.resolution is non-negative, doesn't appear bimodal, and shows right skewness. The Gaussian distribution is symmetrical and allows negative values, which are characteristics that do not match the response variable. The gamma distribution is a better choice because it does not allow negative values and has right skewness. For these reasons, gamma will provide a better model fit and accuracy than Gaussian.

(d) (2 *points*) Discuss general tradeoffs relating to model complexity.

Candidates performed well on this task overall. Full credit was awarded for accurate descriptions connecting model complexity to bias-variance tradeoff or model over- and underfitting.

ANSWER:

As model complexity increases, the model more closely fits the training data. A very simple model (say a constant) is not able to capture any underlying relationships between the target variable and the independent variables. A more complex model may overfit to noise in the training data. Overfitting makes the model less accurate because the model will not perform as well on the testing data. The bias-variance tradeoff is the tradeoff between building a more complex model that detects more patterns in the data and building a less complex model that will generalize better to unseen data.

Your assistant created a GLM. Refer to the code in R.

(e) *(3 points*) Create a new GLM that is a less complex version of the model your assistant made. It should represent the simplest model you can justify using the drop1 output and AIC criteria. Justify whether you would recommend the new model to your boss based on AIC criteria. Copy the code used to create your new model. Copy the output that supports your decision.

This task was relatively straightforward, and candidates performed well overall. Some candidates recognized that the department variable is redundant in a model with type id. However, this observation was not required for full credit.

ANSWER:

Justification:

Eliminating DEPARTMENT is the new model that should be recommended based on AIC. The resulting model is less complex than the original model and does not raise the AIC above the original value.

Code:

```
formula.simpler <- as.formula("Time.to.resolution ~ TYPEid + month + year")

sofa.simpler.glm <- glm(formula.simpler, data = df.sofa, family = Gamma(link ="log"))
```

Output:

```
Single term deletions

Model:
Time.to.resolution ~ DEPARTMENT + area + TYPEid + month + year
           Df Deviance AIC
<none>      3388.0 23189
DEPARTMENT  0 3388.0 23189
area        3 3396.3 23191
TYPEid     10 3532.0 23302
month      11 3690.8 23446
year        1 3690.6 23466
```

Task 4 (*9 points*)

You have worked with your assistant to predict Time.to.resolution for complaints related to mattresses.

(a) (*3 points*) Explain two benefits of a decision tree, focusing on the characteristics of the data itself (refer to the data dictionary).

Candidates performed well on this task overall. Full credit responses discussed two benefits with connection to the dataset.

ANSWER:

One benefit is that decision trees can model non-linear relations between variables without having to specify them in advance. For example, if there is a seasonal relationship between the numeric month variable and the target variable, a decision tree is capable of identifying and modeling that non-linear relationship. By contrast, a GLM would require a transformed version of the month variable in order to reflect a non-linear relationship.

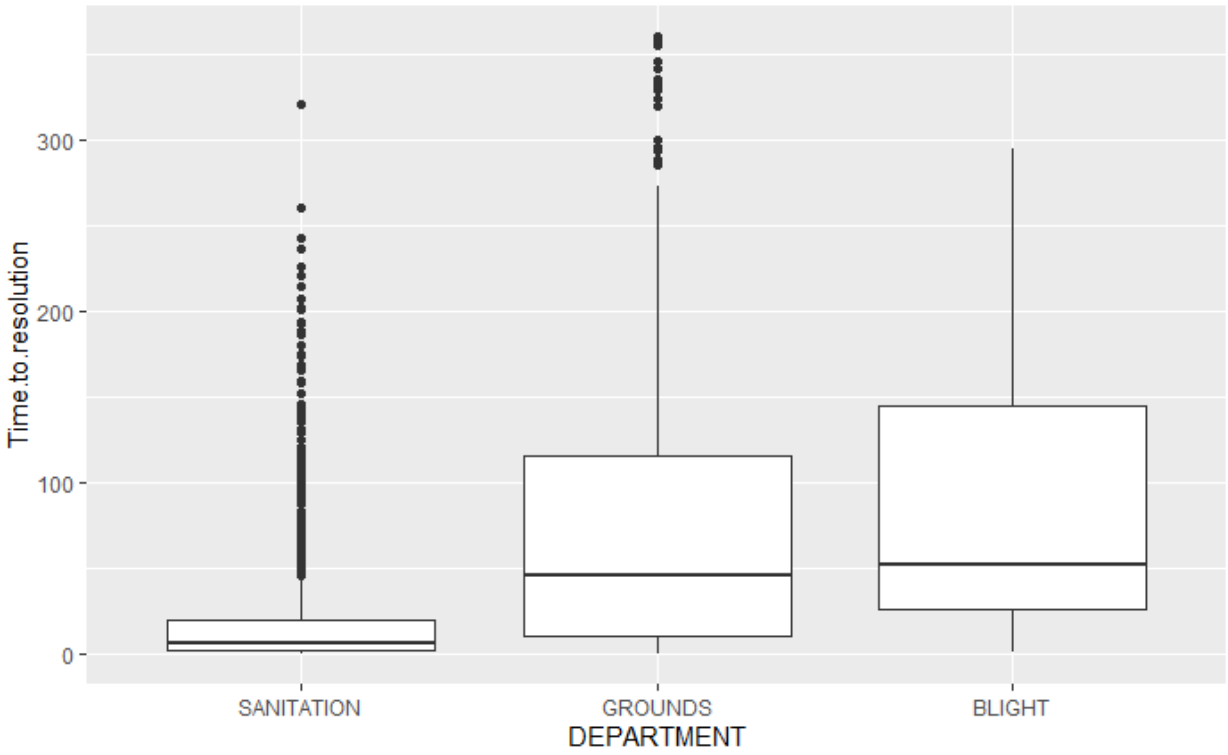
Another benefit is that decision trees can identify interactions. For example, there could be different seasonal patterns depending on geography. Simply including the year, month, latitude, and longitude variables would allow for the tree to model these potential interactions.

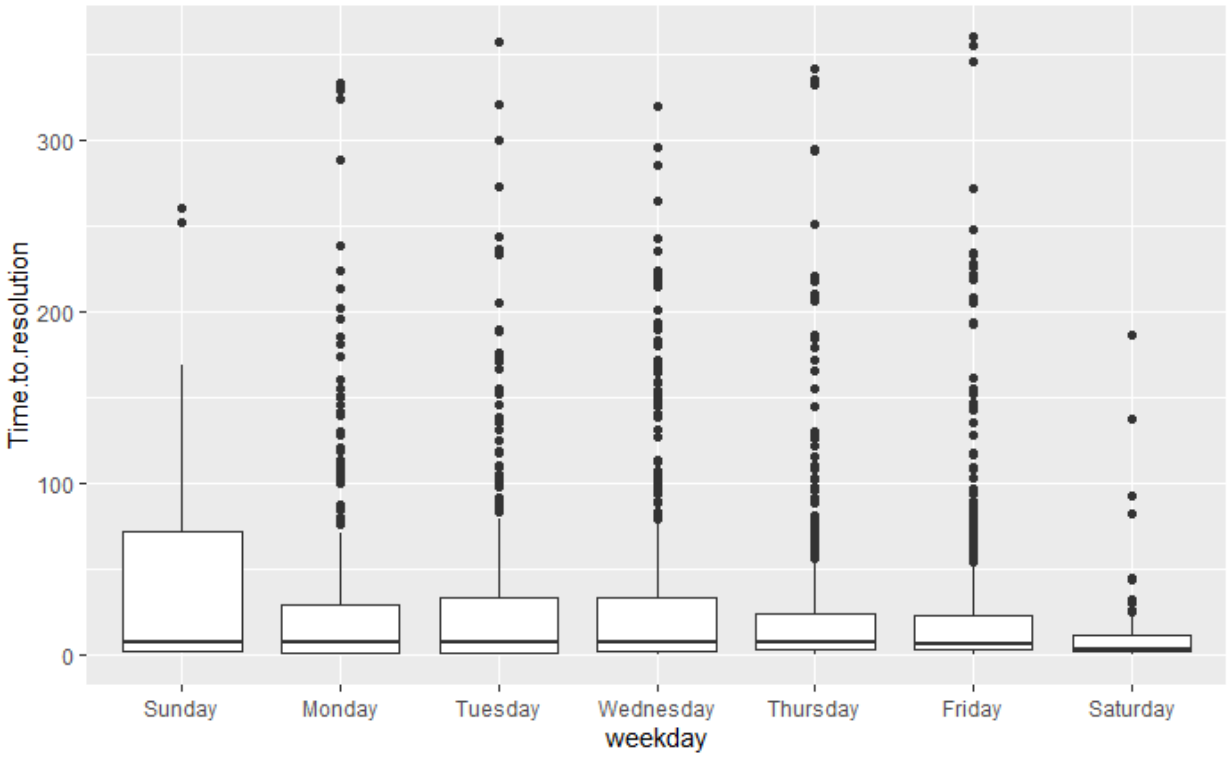
(b) (*3 points*) Assess whether the DEPARTMENT or weekday variable is more likely to be included as an important split in the decision tree. Base your decision on bivariate analyses. Do not create a tree. Paste the results of bivariate exploration that was used to support your assessment.

Candidate performance was mixed on this task. Different types of plots were accepted for full credit provided they were informative in the context of a bivariate analysis.

ANSWER:

Plots:

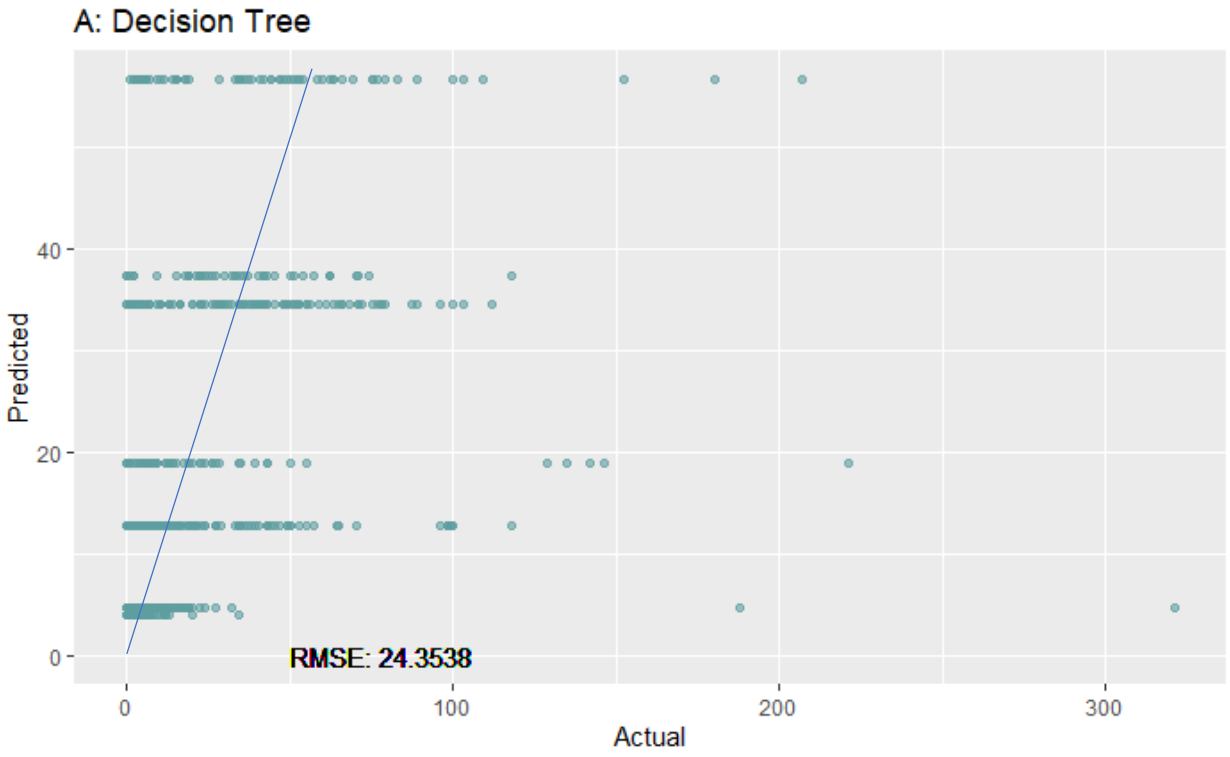


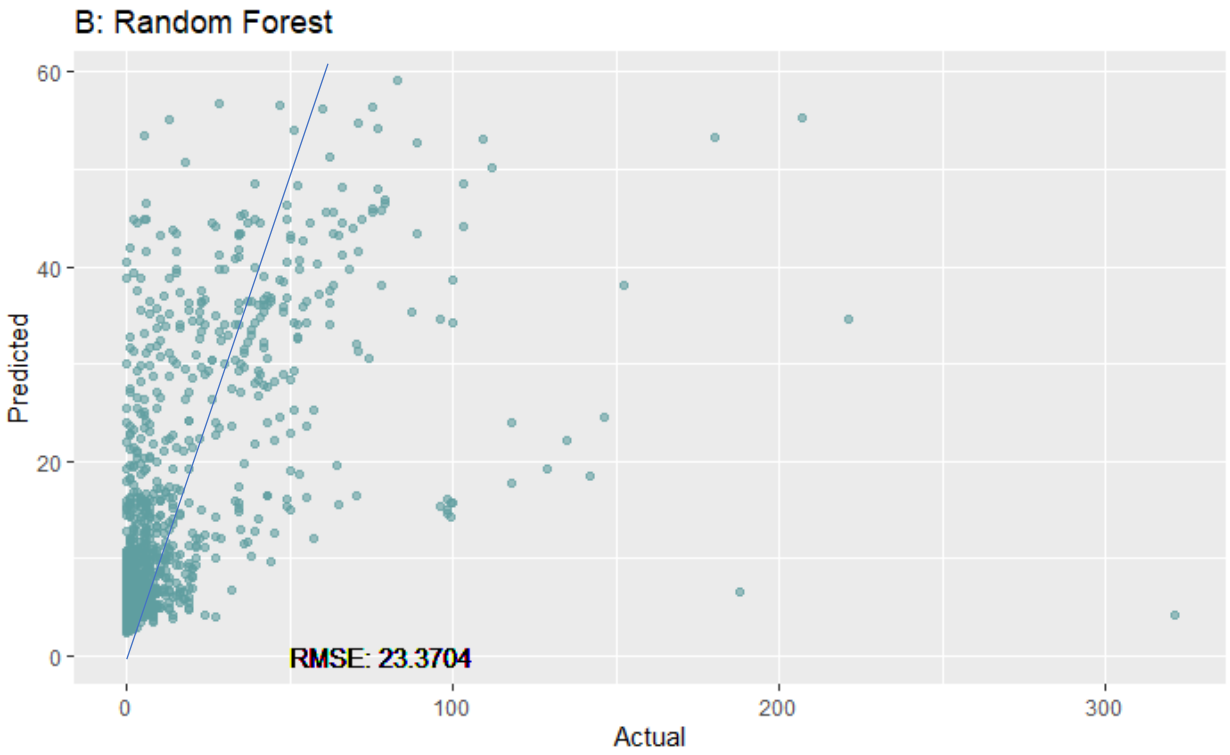


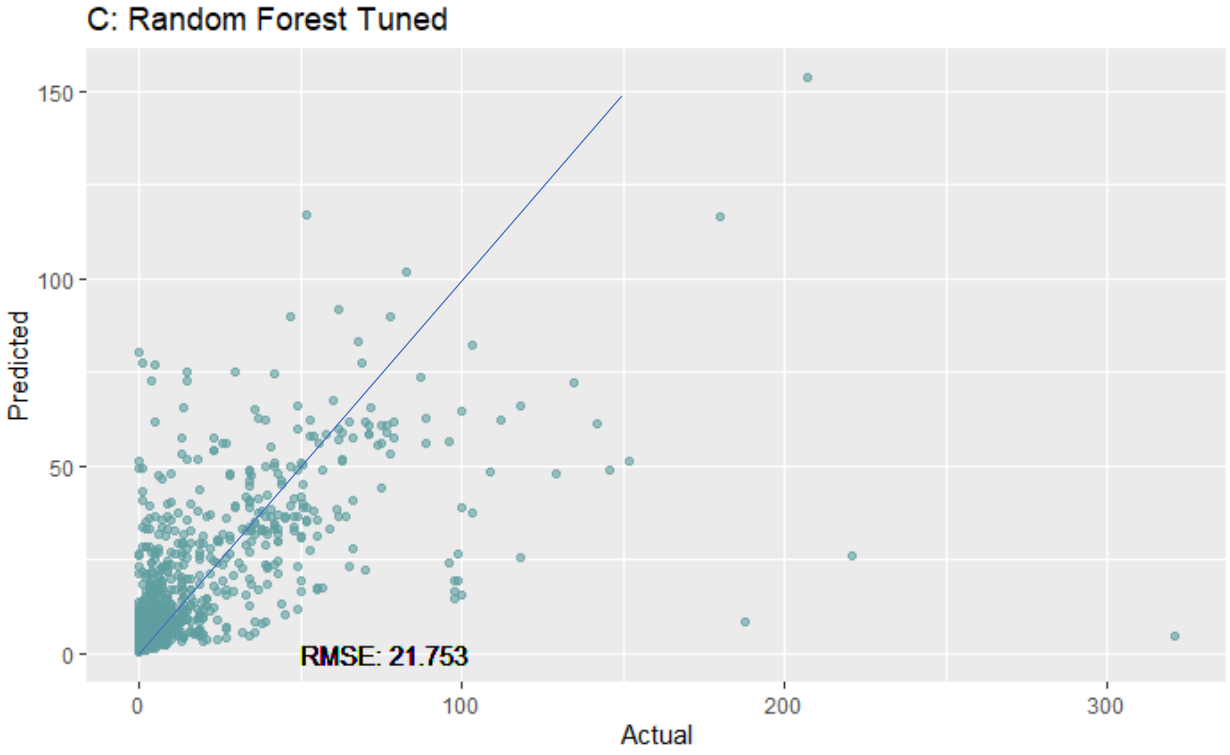
Assessment:

Based on a comparison of the two boxplots above, the department variable looks like it would be more likely than weekday to be a good split. Reviewing the first boxplot, the mean time to resolution for the sanitation department is much lower than for the other two departments (suggesting a potential split in the decision tree). However, when looking at the second boxplot, weekday does not appear to exhibit as much variation between the different days. This leads me to believe it is not likely to be featured as prominently in the decision tree as department.

You ask your assistant to prepare a model to predict Time.to.resolution for mattress complaints coming from the sanitation department. Your assistant produces three different models; a decision tree (A), a random forest with default hyperparameters (B), and a random forest with hyperparameters that have been optimized (C). Plots of the Predicted vs. Actual values of Time.to.resolution for each of the models are below.







(c) (*3 points*) Analyze the differences in the three plots and describe what likely led to those differences.

Candidates performed well on this task overall. Full credit responses correctly identified the differences in the charts and RMSE values as well as making sound inferences into how the corresponding model could generate the differences.

ANSWER:

In plot A, a single decision tree was run. The tree must have had 7 leaf nodes because there are only 7 possible values on the vertical axis.

Plot B is from a random forest, which averages the results of many different trees that are each build on a different subset of observations and variables. This results in many more distinct predicted values.

The RMSE also is shown to have improved from A to B to C. A is by far is the simplest model, and it is not complex enough to detect the signal of the two random forest models. In plot C, a tuning process was used. Random forests are typically "tuned" by changing the number of features used in each split, though other features such as observation sample size, tree complexity, and the number of tree parameters can be adjusted as well. The tuning process typically leads to a more accurate model, which is consistent with the additional drop in RMSE from plot B to plot C.

Task 5 *(5 points)*

Your assistant fit a GLM to predict the resolution time for garbage cart requests from new residents. (The data used is not in any of the supplied files.) The assistant chose to fit two different distributions, a Poisson and a quasi-Poisson distribution. Refer to output below:

```
Call:
glm(formula = Time.to.resolution ~ year + as.factor(month) +
    LONGITUDE + LATITUDE, family = poisson(link = "log"), data = df.task1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.8900  -1.6644  -0.6477   0.4110  30.2298

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       284.909815   5.451537  52.262  < 2e-16 ***
year               -0.140033   0.001607 -87.148  < 2e-16 ***
as.factor(month)2  -0.067987   0.015535  -4.376 1.21e-05 ***
as.factor(month)3   0.156672   0.014400  10.880  < 2e-16 ***
as.factor(month)4   0.065927   0.015191   4.340 1.43e-05 ***
as.factor(month)5   0.193870   0.014664  13.221  < 2e-16 ***
as.factor(month)6   0.091998   0.014512   6.339 2.31e-10 ***
as.factor(month)7   0.022461   0.014757   1.522    0.128
as.factor(month)8   0.794293   0.012942  61.373  < 2e-16 ***
as.factor(month)9   0.282731   0.014287  19.789  < 2e-16 ***
as.factor(month)10  0.695558   0.013310  52.257  < 2e-16 ***
as.factor(month)11  0.261785   0.014945  17.516  < 2e-16 ***
as.factor(month)12 -0.029679   0.016084  -1.845    0.065 .
LONGITUDE           0.104483   0.046603   2.242    0.025 *
LATITUDE            0.304871   0.047265   6.450 1.12e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 125513  on 14324  degrees of freedom
Residual deviance: 104664  on 14310  degrees of freedom
AIC: 158466

Number of Fisher Scoring iterations: 6
```

```
Call:
glm(formula = Time.to.resolution ~ year + as.factor(month) +
    LONGITUDE + LATITUDE, family = quasipoisson(link = "log"),
    data = df.task1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.8900  -1.6644  -0.6477   0.4110  30.2298

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       284.90981   19.81415  14.379  < 2e-16 ***
year               -0.14003    0.00584 -23.977  < 2e-16 ***
as.factor(month)2  -0.06799    0.05646  -1.204 0.228581
as.factor(month)3   0.15667    0.05234   2.994 0.002763 **
as.factor(month)4   0.06593    0.05521   1.194 0.232483
as.factor(month)5   0.19387    0.05330   3.638 0.000276 ***
as.factor(month)6   0.09200    0.05275   1.744 0.081156 .
as.factor(month)7   0.02246    0.05363   0.419 0.675384
as.factor(month)8   0.79429    0.04704  16.886  < 2e-16 ***
as.factor(month)9   0.28273    0.05193   5.445 5.28e-08 ***
as.factor(month)10  0.69556    0.04838  14.378  < 2e-16 ***
as.factor(month)11  0.26178    0.05432   4.819 1.45e-06 ***
as.factor(month)12 -0.02968    0.05846  -0.508 0.611686
LONGITUDE           0.10448    0.16938   0.617 0.537349
LATITUDE            0.30487    0.17179   1.775 0.075974 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 13.21031)

    Null deviance: 125513  on 14324  degrees of freedom
Residual deviance: 104664  on 14310  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

(a) (3 *points*) Assess the two chosen distributions with respect to reasonability in modeling Time.to.resolution as a target variable, using the output provided by the assistant.

Few candidates received full credit on this task. Full credit responses identified overdispersion from the model output and explained its implications in interpreting and using model output from models using the respective distributions. Partial credit was awarded for responses that identified numerical differences between the output of the two models but did not connect the differences to properties of the two distributions. Minimal partial credit was awarded to candidates who remarked only on the applicability of the Poisson distribution without contrasting it with the quasi-Poisson.

ANSWER:

An underlying assumption of Poisson regression is that the mean and variance are equal. The assistant's code shows that the variance is greater than the mean for new resident requests, indicating evidence of overdispersion.

Quasi-Poisson regression is equipped to deal with the problem of overdispersion. Notably, the estimates of the coefficients are the same when compared to the Poisson output. However, the standard errors are all higher and fewer coefficients are statistically significant (coefficients for months 2, 4, 7, 12, and LONGITUDE are not significant in the quasi-Poisson output, whereas only month 7 is not significant in the Poisson output).

While both distributions could be used for modeling as they ultimately lead to the same predictions, if any further analysis is conducted such as deriving confidence intervals or conducting hypothesis tests, the quasi-Poisson distribution should be used.

Your boss would like you to consider other distributions for the GLM.

(b) (*2 points*) Recommend two additional distributions along with link functions that are reasonable choices to model Time.to.resolution. Justify your recommendations.

Candidates performed well on this task overall. Full credit responses recommended distributions and link functions with justification based on the characteristics of the target variable such as the domain and skewness of the data. A common reason for candidates receiving only partial credit was justifying only the distribution or the link function.

ANSWER:

I recommend fitting the following, with the target variable being Time.to.resolution + 1:

(1) Gamma distribution with a log link function
(2) Inverse Gaussian with a log link function

Adding 1 (or another small positive value) to Time.to.resolution is necessary since the gamma and inverse Gaussian distributions do not support 0 values, which do exist in the Time.to.resolution variable. Using the log link function ensures that the predictions are positive.

The two recommended distributions support continuous values while the target variable contains only integer values. However, these distributions are still practical choices given the target variable is positive and right-skewed.

Task 6 *(10 points)*

The client is interested in improving furniture disposal pickup times. Your assistant prepares a GLM and a decision tree that model Time.to.resolution using LATITUDE and LONGITUDE as predictor variables. (The data used is not in any of the supplied files.)

(a) (*2 points*) Contrast using a GLM versus a decision tree given the client's goals and the variables chosen to use in these models.

Note: There was overlap in rationale candidates provided in task 6(a) and 6(b). Credit was awarded for reasonable justifications whether they were provided as a response to 6(a) or 6(b). This commentary applies to both 6(a) and 6(b).

Performance was mixed on these tasks. Full credit responses described how each model would interpret the geospatial data and contrasted those observations in the context of the client's goal to treat neighborhoods equally. Most full credit responses identified a weakness in using a GLM based on the raw geospatial variables and how a decision tree can overcome the weakness. Credit was awarded for discussing either why the linear relationship assumption or the inability of a GLM to identify interactions is unreasonable. Partial credit was awarded to responses that addressed these weaknesses in vague terms, e.g., stating that decision trees have more flexibility to fit to a variable.

ANSWER:

A GLM trained on LATITUDE and LONGITUDE as predictor variables will produce coefficients representing how the target variable (time.to.resolution) varies linearly from west to east and from south to north across the city. If time.to.resolution does vary geographically, it seems more likely that there will be certain sections of the city (e.g., neighborhoods, districts) where the time.to.resolution may be higher or lower than the city-wide average, as compared to a linear function of the time.to.resolution across the city.

A decision tree will be more adept at identifying sections of the city with higher or lower times to resolution compared to the city wide average by determining several splits in which the city may be divided into any number of rectangular sections. Also, decision trees do not assume that the target variable has linear dependence on the predictor variables.

(b) (*2 points*) Recommend either the GLM or decision tree to use and justify your recommendation.

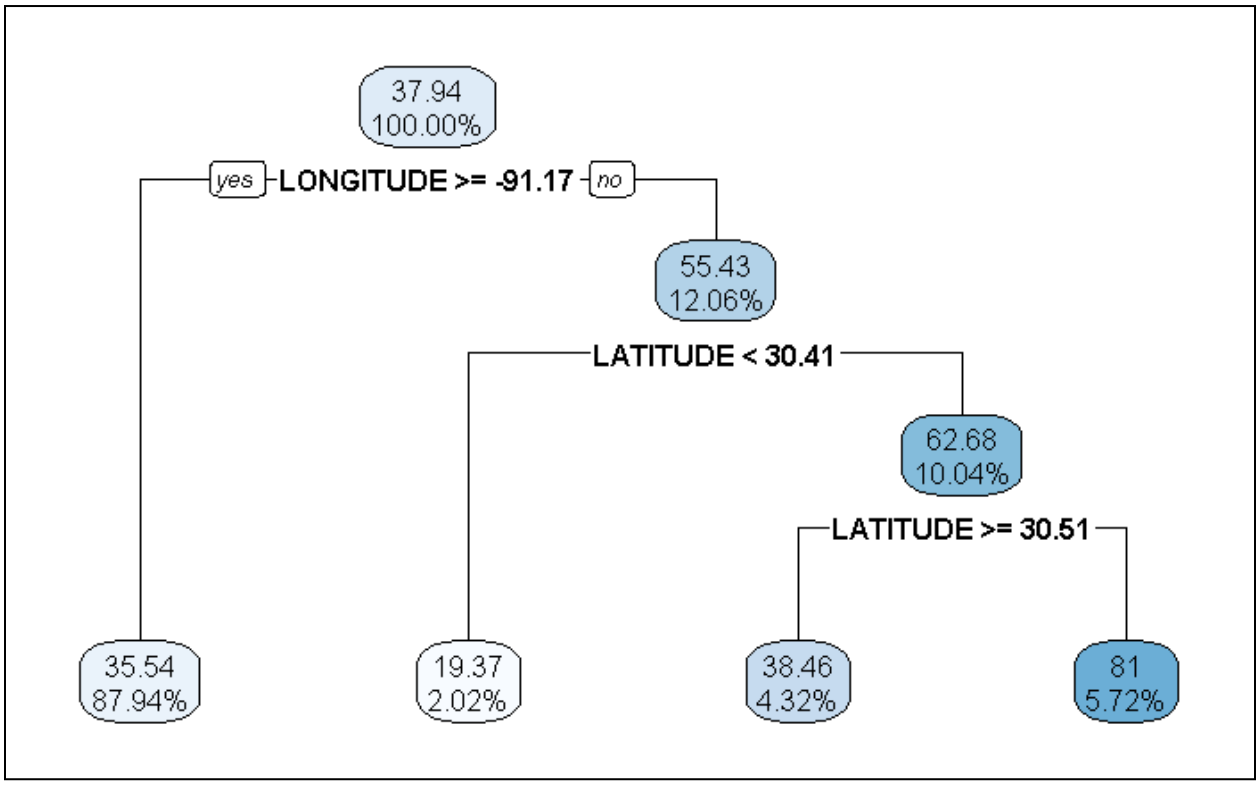
See note for 6(a).

ANSWER:

A decision tree will produce splits that would allow the client to divide the city into areas with longer and shorter times to resolution. The decision tree also does not assume linear dependence on the predictor variables like the GLM does and there is no reason to expect that a linear relationship should exist. For these reasons, I recommend using the decision tree for this problem.

Your assistant produced a decision tree to predict Time.to.resolution using LATITUDE and LONGITUDE. Your assistant provides you with the following code and output below:

```
formula <- as.formula("Time.to.resolution~LATITUDE+LONGITUDE")
tree.furniture <- rpart(formula,data=df.furniture,cp=.003,minbucket=50)
rpart.plot(tree.furniture,type=2,digits=4)
```



(c) (*3 points*) Interpret a few select components of this plot by filling out the table below:

Candidates performed well on this task overall. Most candidates were able to interpret the decision tree with either no errors or only minor errors.

ANSWER:

Component of Plot	Interpretation
55.43	The average time.to.resolution for all records in the training data where the LONGITUDE is *not* greater than or equal to -91.17
12.06%	The percent of all records in the training data where the LONGITUDE is not greater than or equal to -91.17.
Latitude < 30.41	This is the split that produces the highest reduction of node impurity for all records of the training data that have a LONGITUDE not greater than or equal to -91.17. In other words, within all training data records with longitude less than -91.17, further splitting the data using LATITUDE less than 30.41 as a splitting criteria produces the greatest difference in average time.to.resolution of the resulting leaves, while also requiring that hyper

	parameter pruning measures are met (e.g. minimum number of records in leaves). Records with LATITUDE less than 30.41 are grouped in the left leaf from this node, and records with LATITUDE greater than or equal to 30.41 are grouped in the right leaf from this node.
38.46	This is the average time.to.resolution for all records that end up in this leaf. Records in this leaf are those with LONGITUDE less than -91.17, and LATITUDE greater than or equal to 30.51.
5.72%	This the percent of all records that end up in this leaf. Records in this leaf are those with LONGITUDE less than -91.17, and LATITUDE between 30.41 and 30.51

Your assistant wants to recommend that the client includes shortening furniture disposal service request resolution times as part of their campaign.

(d) (*3 points*) Critique your assistant's recommendation and consider model efficacy and potential equity concerns.

Candidates struggled with this task. Full credit responses addressed both equity and efficacy. The most common model efficacy concern that received full credit was that the model is only built on geospatial variables, and other variables should be considered to add predictive power. Similarly, the most common equity concern was that geospatial variables could be a proxy for protected classes.

ANSWER:

Based on the plot of the decision tree in subtask b, the time.to.resolution for service requests relating to furniture varies significantly by geography. The assistant's recommendation to the client seems reasonable based on the results of the decision tree. However, there are several considerations that should be discussed with the client before they move forward with using this recommendation. A few of these considerations are listed as follows:

- Allocating resources geographically could have negative downstream effects if not done carefully. Race, ethnicity, socio-economic status, age, and other demographics often vary geographically. Focusing garbage collection resources in some districts over others could indirectly produce differences in resource access for one of these demographic categories listed, which could have negative political or social impacts on the client.
- The decision tree supporting this decision is fairly simple and does not contain details regarding the distribution of time until resolution by node. For example, if a couple of outliers are contributing towards the node with the long time until resolution, this could skew the interpretation of this plot and potentially provide the client with a false focus.
- This decision tree only focuses on LATITUDE and LONGITUDE to predict the time until resolution. There may be other strong predictors as well. It might be wise to have a more complete picture of the predictors before reaching out to the client with recommendations.

Task 7 (*8 points*)

Your assistant is interested in better understanding debris removal on unoccupied property and wants to use hierarchical clustering to separate the data into subgroups. Your assistant prepares a dataset that contains debris removal records on unoccupied property on or after 2019. (The data used is not in any of the supplied files.)

Consider a dataset with three variables X, Y, and Z, and three observations A, B, and C. Suppose they contain the following observations:

	X	Y	Z
Observation A	1	2	1
Observation B	2	1	2
Observation C	10	20	10

(a) *(2 points)* Identify which two observations are closest together using each of the dissimilarity measures of correlation and Euclidian distance.

Candidates performed well on this task overall. Candidates were not required to perform any calculations if they provided sound rationale to deduce their answer.

ANSWER:

Using Euclidean distance, observations A and B will be in closest proximity because the squared difference on X, Y, and Z is each 1, whereas the distance from either point to C is at least 8^2 = 64 for variables X, Y, and Z.

Dissimilarity is measured using correlation. Since each variable for observation C is exactly 10 times observation A, these two observations will be closer together using correlation as a dissimilarity measure.

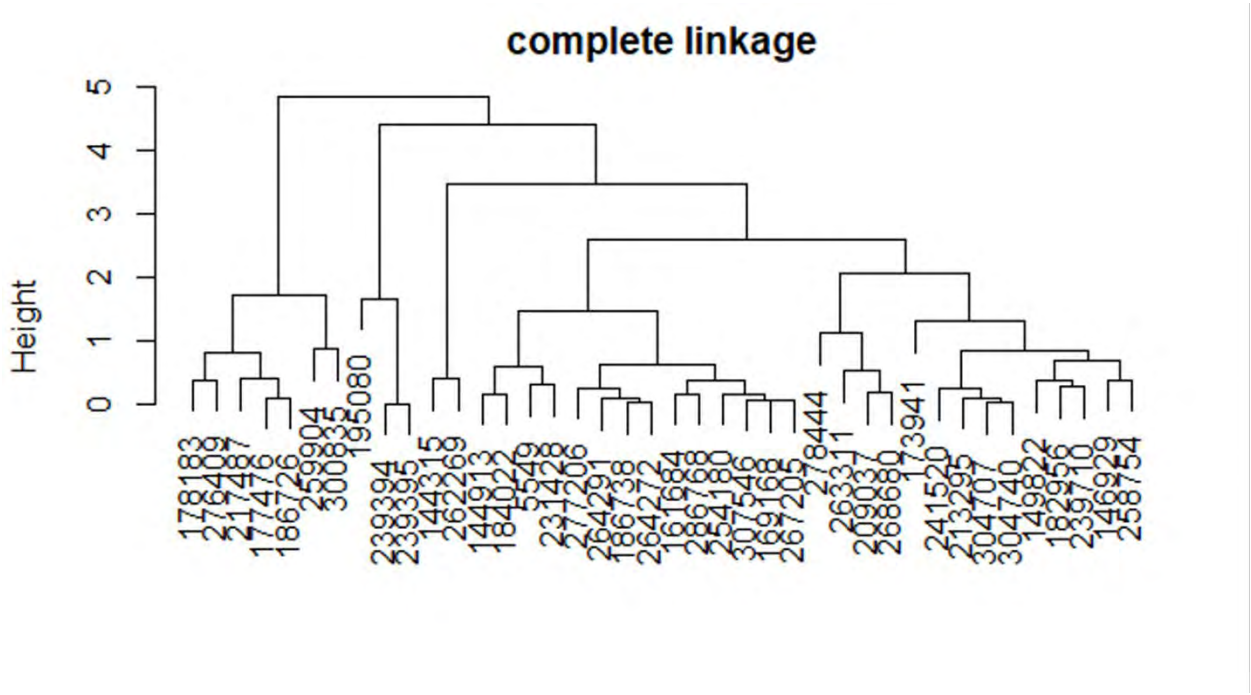
(b) (*1 point*) State the difference between dissimilarity and linkage.

Candidate performance was mixed on this recall question.

ANSWER:

Dissimilarity measures the proximity of two observations in the data set, while linkage measures the proximity of two clusters of observations.

Your assistant performs hierarchical clustering with the variables LONGITUDE and LATITUDE, using the complete linkage method.



(c) (*2 points*) Interpret what the height represents in the complete linkage dendrogram above.

Overall candidate performance was poor on this task. Full credit responses demonstrated an understanding of what height means in a dendrogram and the implications of splits occurring at different heights.

ANSWER:

The height values correspond to the values of the linkage function at the points where two clusters are combined in the hierarchical clustering algorithm.

When the height difference between adjacent values is small on a relative basis, the observations in the corresponding clusters have similar characteristics as determined by the linkage function and dissimilarity measure, and it makes sense to combine the clusters.

On the other hand, when the height difference between adjacent values is large, the observations in the corresponding clusters have materially different characteristics, so combining the clusters would result in a loss of information.

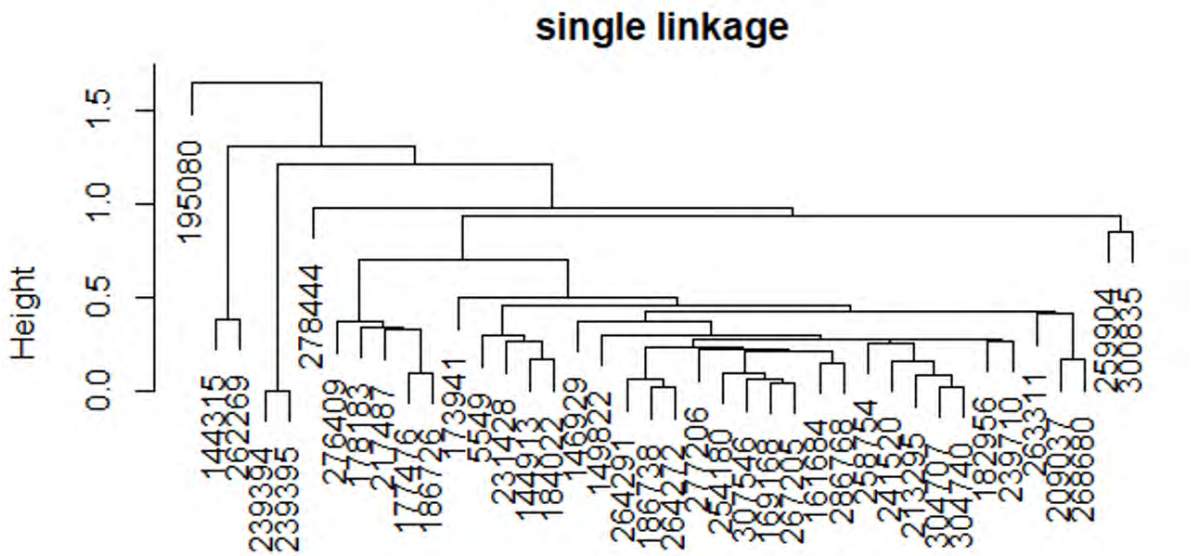
(d) (*1 point*) Recommend a value for the number of clusters for the complete linkage dendrogram. Justify your recommendation.

Candidates performed well on this task overall. Full credit was awarded for any well-justified recommendation.

ANSWER:

I recommend 4 clusters. 5 clusters would make the clustering more balanced, but there isn't a very large difference in height between the additional cluster that would break off and its parent.

Your assistant runs a similar clustering algorithm, this time with single linkage.



(e) (2 *points*) Explain how the dendrogram from the single linkage method differs from the complete linkage dendrogram.

Candidate performance was mixed on this task. Full credit responses demonstrated understanding of how the clustering algorithm works and why the complete method tends to produce clusters that are more balanced.

ANSWER:

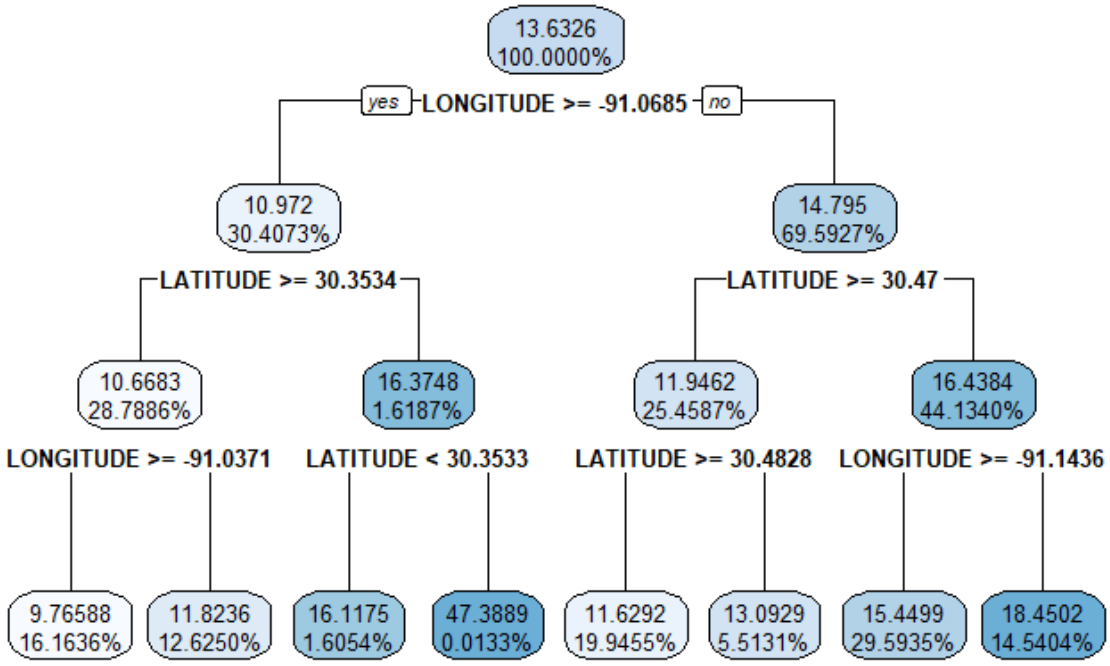
The clustering algorithm starts out with n clusters and fuses them together in an iterative process based on which observations are most similar. The complete linkage method considers the maximum intercluster dissimilarity and the single linkage method uses minimum intercluster dissimilarity. As such, the single method tends to fuse observations on one at a time resulting in less balanced clusters.

Task 8 (*7 points*)

Your assistant prepared two decision trees. Each tree is trained on a subset of service requests related to missed garbage pickups. The two trees are similar except for the target variable. The code and plots of these trees are shown as follows, which are used in subtasks (a), (b), and (c). (The data used is not in any of the supplied files.)

tree.1:

```
formula.tree.1 <- as.formula("Time.to.resolution~LATITUDE+LONGITUDE")
tree.1 <- rpart(formula.tree.1,data=df.missed,cp=0,maxdepth=3)
rpart.plot(tree.1,type=2,digits=6)
```

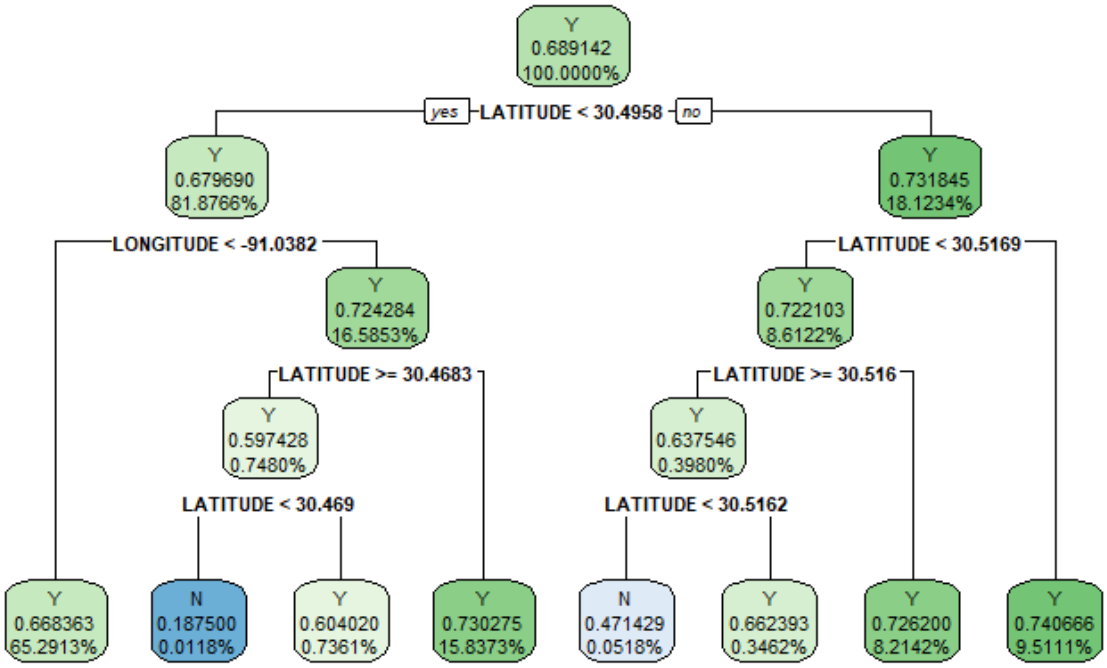


tree.2:

```
df.missed$res.under.ten.days[df.missed$Time.to.resolution < 10]<-"Y"
df.missed$res.under.ten.days[df.missed$Time.to.resolution >= 10]<-"N"
df.missed$res.under.ten.days <- as.factor(df.missed$res.under.ten.days)

formula.tree.2 <- as.formula("res.under.ten.days~LATITUDE+LONGITUDE")
tree.2 <- rpart(formula.tree.2,data=df.missed,cp=0,maxdepth=4)

rpart.plot(tree.2,type=2,digits=6)
```



(a) (3 *points*) Explain how the calculations to determine the splits in tree.1 are different than in tree.2.

Candidate performance was mixed on this task, with most candidates receiving varying amounts of partial credit. Full credit responses correctly identified tree.1 as regression and tree.2 as a classification tree and correctly explained the calculation mechanics for both types of trees.

ANSWER:

Tree.1 is a regression decision tree because Time.to.resolution is a continuous numeric variable and Tree.2 is a classification tree because res.under.ten.days is a binary classification variable.

Regression trees determine splits by first measuring the residual sum of squares errors between the target and predicted values. The splits are generated by maximizing the difference between the nodes and minimizing the difference within the nodes.

Classification decision trees measure impurity using entropy, Gini index, or classification error. These measures all attempt to increase the homogeneity of the target variable at each split.

Your boss wants to understand the impact of outliers in the models.

(b) *(2 points)* Explain how the presence of outliers in the target variable affects each of the two trees from part (a) above.

Candidates performed well on this task overall. Full credit was awarded to candidates who recognized that the regression tree is more susceptible to outliers and provided a clear explanation.

ANSWER:

The regression and classification trees will likely have similar patterns in their respective splits, but may not be identical. For example, if there are outliers in the Time.to.resolution variable, splits in the decision tree may favor the outlier due to the additional leverage. However, the res.under.ten.days variable is less sensitive to outliers because all values above 10 are treated the same, regardless how extreme they may be.

Your assistant prepares a draft version of the report to the client that includes both trees. To simplify the report, you decide to only include one of the trees.

(c) (2 *points*) Recommend a tree to present to the client. Justify your recommendation based on the applicability to the business problem.

Candidates performed well on this task overall. Credit was awarded for either recommendation provided it had sound justification.

ANSWER:

I recommend sending the client the regression tree (tree.1) and not the classification tree (tree.2). Both trees are similar, but the regression tree is more directly applicable to the business problem of "predicting time to resolution", which may be more useful in the client's campaign. A classification tree that measures 10-day pickup is not as directly useful for the business problem.

Task 9 *(6 points)*

The client is interested in estimating the impact of various predictors on Time.to.resolution for two common complaints: "MISSED GARBAGE SERVICE DAY (GENERAL PICK-UP)" and "MISSING GARBAGE CART." The client is interested in resolution time trends. Another concern is whether resolution times differ for certain areas within the city.

Run the given code and use the output to answer the following.

(a) (*3 points*) Interpret the coefficients for the time variables (year, quarter) for the two models (one for each complaint) using the summary() output. Also describe the trends of resolution times for each of the two complaints.

Candidates performed well on this task overall. Full credit responses interpreted coefficients in the context of the model form, interpreted the coefficient signs, and compared the coefficient trends across models.

ANSWER:

Model 1: Missed Garbage Service Day (General Pick-up)

Variable	**Coefficient**	**Exp ()**	**Decreasing by**	**Interpretation**
Year	-0.82764	0.43709	0.56292 (56%)	Each year, missed general pickup is expected to decrease by 56%, all other variables held equal
Q2	-0.15566	0.85585	0.14415 (14%)	Q2 is expected to have 14% fewer missed pickups than Q1, all other variables held equal
Q3	-0.74749	0.47355	0.52645 (53%)	Q3 is expected to have 53% fewer missed garbage carts than Q1, all other variables held equal
Q4	-0.83595	0.43346	0.56654 (57%)	Q4 is expected to have 57% fewer missed garbage carts than Q1, all other variables held equal

Model 2: Missing Garbage Cart

Variable	**Coefficient**	**Exp ()**	**Change**	**Interpretation**
Year	-0.17612	0.83852	0.16148 (-16%)	Each year, missing garbage carts is expected to decrease by 16%, all other variables held equal
Q2	0.25359	1.28865	0.28865 (+29%)	Q2 is expected to have 29% more missed garbage carts than Q1, all other variables held equal

Q3	0.34085	1.40615	0.40615 (+41%)	Q3 is expected to have 41% more missed garbage carts than Q1, all other variables held equal
Q4	0.38793	1.47393	0.47393 (+47%)	Q4 is expected to have 48% more missed garbage carts than Q1, all other variables held equal

The year coefficients are negative for both complaint TYPEs. This implies that response times are reducing over time, from year to year. For the "missed general pickup" TYPE, the time.to.resolution improves as you move into later quarters in a year. For the "missing garbage cart" we see the opposite trend with time.to.resolution increasing as you move into later quarters in a year.

(b) (*3 points*) Using the summary and drop1 output, compare and contrast the significance of the area variables in the two models. Quantify significant differences in resolution times.

Candidates performed well on this task overall. Full credit responses discussed the output of the summary and drop1 functions, provided justification for why variables were significant or not, and quantified significant differences in response time over the baseline.

ANSWER:

Based on summary output:

- In model 1, the area variable is significant with a p-value of 3.02e-08. The coefficient for areaW = 0.20493 means that time.to.resolution for areaW is higher than the reference level, areaD, by exp(0.20493) – 1 = 22.74%.
- In model 2, none of the area categories are significant with all p-values greater than 0.5.

Based on drop1 output:

- In model 1, dropping the area variable gives a higher AIC, suggesting that the area variable should not be dropped.
- In model 2, dropping the area variable gives a lower AIC, suggesting that the area variable should be dropped.

Task 10 *(10 points)*

Your client has a goal to resolve missed pickups service calls to fewer than two days. Your boss wants you to build a model to evaluate this and suggests using AUC as a performance metric. (The data used is not in any of the supplied files.)

(a) *(2 points)* Explain the difference between accuracy and AUC in terms of overall model assessment.

Candidate performance was mixed on this task. Full credit responses provided correct definitions of both accuracy and AUC and recognized the key difference that accuracy is calculated at a single cutoff point while AUC is calculated across all possible cutoff points. Most candidate responses failed to provide a complete response.

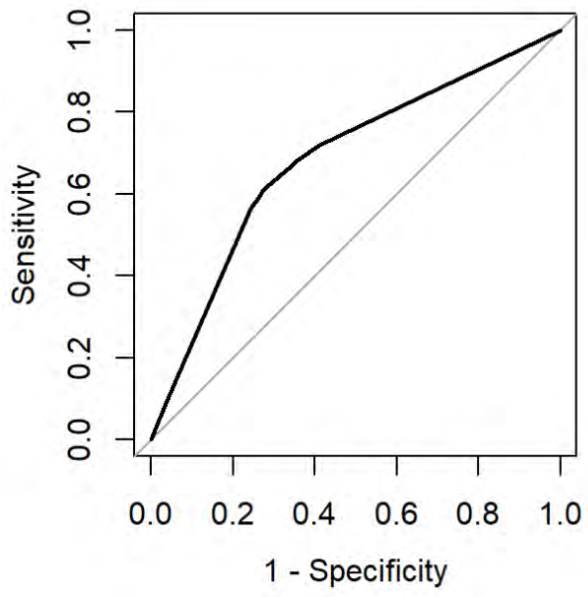
ANSWER:

Accuracy is measured by the ratio of correct number of predictions to total number of predictions made. It doesn't directly utilize the modeled probabilities, but rather the classifications based on a fixed cutoff point.

AUC measures the area under the ROC curve. It assesses the overall model performance by measuring how true positive rate (TPR) and false positive rate (FPR) trade off across a range of possible classification thresholds.

AUC measures performance across the full range of thresholds while accuracy measures performance only at the selected threshold.

Your assistant built a model and plotted the ROC curve below.



(b) (2 points) Explain why the ROC curve always goes through (0,0) and (1,1).

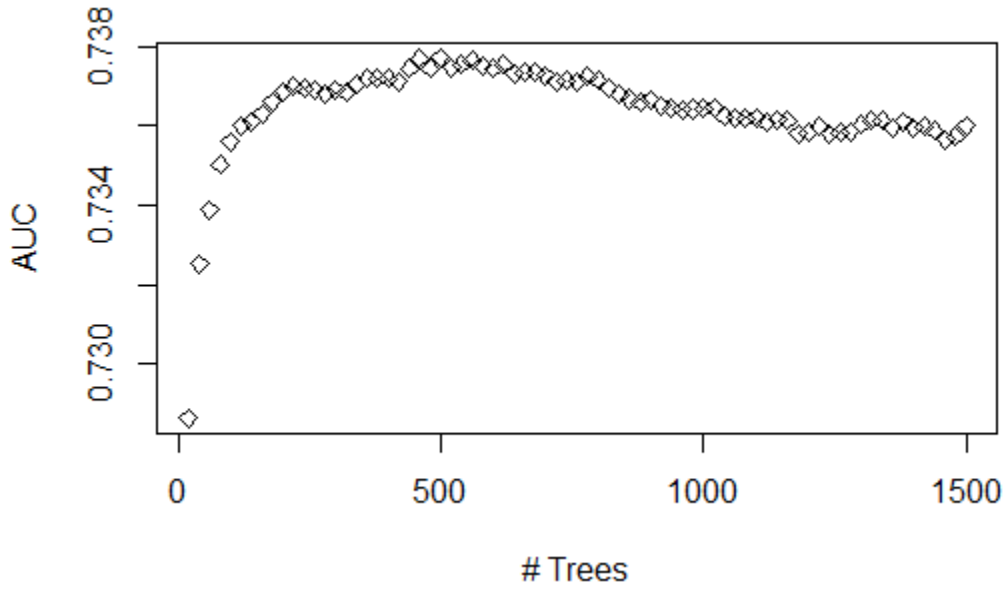
Candidate performance was mixed on this task. Common errors were not mentioning how points on the ROC curve vary by threshold and incorrectly defining Sensitivity or Specificity.

ANSWER:

The ROC curve plots Sensitivity on the y-axis and (1-Specificity) on the x-axis. The point (0,0) corresponds to a Sensitivity value of 0 and a Specificity value of 1, meaning 1-Specificity equals 0. Setting our classifier so that it never identifies a positive case will produce (0,0) because the true positive rate (Sensitivity) will be zero and with every case classified as negative the true negative rate will be 1. The (1,1) point is the opposite case where everything is classified as positive and the true positive rate (Sensitivity) rises to 1, while the true negative rate (Specificity) drops to 0.

Your boss suggests a boosted tree can increase model performance by reducing bias, however, setting hyperparameters is critical. You are asked to build a gradient boosting machine (GBM) tree model to assess the performance improvement.

The GBM tree model performance using the test data set is shown below.



(c) (2 *points*) Explain why model performance improves at beginning then deteriorates as the number of trees increases.

Candidate performance was mixed on this task. Full credit responses identified that the pattern of performance on the test data results from overfitting and provided enough description of the mechanics of a GBM to explain how the overfitting occurred.

ANSWER:

A GBM iteratively builds trees fit to the residuals of prior trees. Depending on the hyperparameters, this model can produce a very complex model which is susceptible to overfitting to patterns in the training data.

In this model, AUC on the testing data increases until the number of trees reaches about 500. However, as more trees are added beyond 500, AUC on the testing data starts to drop, which indicates the model is overfit to the training data.

(d) (2 *points*) Describe two hyperparameters you could adjust to improve model performance.

This task was straightforward, and candidates performed well overall. Any GBM hyperparameters could be chosen provided a correct justification.

ANSWER:

Early stopping: Early stopping criteria, such as improvement of the performance metrics in each subsequent tree, can stop training when it detects the improvement is marginal. This avoids overfitting.

Controlling learning rate: Learning rate controls the impact of subsequent trees to the overall model outcome. This reduces the extent to which a single tree is able to influence the model fitting process.

(e) (*2 points*) Explain the process of how to tune a hyperparameter.

Candidate performance was mixed on this task. Full credit responses described cross-validation and how it is used in the hyperparameter tuning process. Many candidates provided overly vague responses, often with no reference to cross-validation.

ANSWER:

Tuning a hyperparameter requires first varying the hyperparameter across a range of possible values and performing cross validation at each value. Performance is then determined based on a cross-validation performance metric, for example AUC, and the hyperparameter value with best performance based on this metric is selected.

Task 11 *(8 points)*

You are investigating data on calls for damaged carts using Time.to.resolution as the target variable. This dataset includes an additional variable "Service.Request.Id." This variable is set to 1 for the first request and incremented by one at each subsequent request. Your assistant has removed this variable, arguing that it is not of any value for predicting Time.to.resolution, given that is merely a counter that reflects the row of the observation. (The data used is not in any of the supplied files.)

(a) (2 *points*) Critique the assistant's recommendation.

Most candidates received partial credit on this task. Full credit responses stated an opinion of the assistant's recommendation and justified it with a thoughtful discussion of how the variable could be used in modeling.

ANSWER:

I disagree with the assistant's recommendation as stated. This variable should not be removed without further investigation. Changes in the ID values over time could identify useful information such as changes in systems or changes in data collection approaches.

Before removing or eliminating a predictor you should at least check for any correlation between it and the target or other predictor variables. It is best to check for patterns or other characteristics that may be of use in feature generation.

You may need to consider multicollinearity with other date and time variables. Multicollinearity could cause issues with many model fitting algorithms.

(b) (*1 point*) Define an interaction effect.

Candidates generally performed well on this recall question.

ANSWER:

An interaction effect is when the target variable has a relationship with a combination of input variables in addition to potentially having a relationship with those variables on their own.

Many service calls for damaged carts have resolution times over 60 days. You have been asked to look at these in more detail. Your assistant has built an initial model to predict if a damaged cart call will take more than 60 days to service. The predictive variables used are: year, month, DEPARTMENT, LATITUDE, LONGITUDE. Consider interactions among the predictor variables.

(c) (*2 points*) Propose two variables to make an interaction term that may improve model accuracy. Justify your proposal.

Candidates performed well on this task overall. The most common full credit responses proposed an interaction between department and year. However, some candidates received full credit for proposing an interaction between LATITUDE and LONGITUDE.

ANSWER:

An interaction term between year and DEPARTMENT could improve the model fit. The various departments may trend differently over time and this nuance can be captured through a time variable and department interaction.

You continue working on a model to predict if a call for a damaged cart will have resolution times over 60 days. A new indicator variable "Over60" has been created to identify records that have a resolution time greater than 60 days.

Your assistant is testing different link functions for predicting Over60. Your assistant notes that some model predictors are highly statistically significant with certain link functions but not with others.

(d) (*3 points*) Explain how changing the link function in the GLM impacts the model fitting and how this can impact predictor significance.

Candidates struggled with this task overall. Full credit responses defined what a link function is, explained how link function impacts error terms, which ultimately impacts model fitting and whether predictor variables are statistically significant.

ANSWER:

The link function specifies a functional relationship between the linear predictor and the mean of the distribution of the outcome conditional on the predictor variables. Different link functions have different shapes and can therefore fit to different nonlinear relationships between the predictors and the target variable. For example: if predictor variables have very linear relationships to the mean, a link function that preserves that linearity (like the identity link function g(u) = u) should provide a better model fit than a link function that creates a more nonlinear, curved relationship to the mean.

When the link function matches the relationship of a predictor variable, the mean of the outcome distribution (the prediction) will generally be closer to the actual values for the target variable, resulting in smaller residuals and more significant p-values.

Task 12 *(7 points)*

Your assistant mentions that using latitude and longitude for each service call would allow the mapping of each call to a zip code. By using publicly available census information, the data by zip code could be combined with information such as average age, predominant race, and average household income.

(a) (*1 point*) Define proxy variable.

Candidate performance was mixed on this task. Responses that demonstrated understanding that proxy variables provide information about a variable that was not directly measured received full credit. A common response that received no credit was defining a proxy variable as a variable that contains sensitive information.

ANSWER:

Proxy variables are variables that are used in place of other information, usually because the desired information is either impossible or impractical to measure. For a variable to be a good proxy it must have a close relationship with the variable of interest.

(b) (3 *points*) Evaluate your assistant's recommendation for any potential legal or ethical concerns including whether proxy variables should be used in this project.

Candidate performance was mixed on this task. Full credit responses Included a description of legal/ethical issues, a discussion of whether there is problematic information in the data that should be addressed, and additional justification based on the context of the business problem.

ANSWER:

Data such as race, age, and income are generally considered sensitive information. Some jurisdictions have legal constraints on the use of sensitive information. Before proceeding there should be consideration of any applicable law. There are no clear rules for what ethical use of data is. Good professional judgement must be used to ensure that inappropriate discrimination is not occurring within the model or the project. Public perception should also be considered. The politician or the city could suffer bad press if there is a belief that the project inappropriately discriminates.

Since LATITUDE and LONGITUDE were used to lookup race, age, and income, including them in a model is clearly creating proxy variables for this sensitive information. Simply not using the census data may not be completely safe for eliminating potential inappropriate discrimination in the model or the project.

However, given the project goal of identifying inequities, using the census data may be a valid way to test that the model and project decisions are fair or even improve equity among demographic groups.

Your assistant states that the values for latitude and longitude are too granular and proposes that the data be grouped for modeling. Your assistant groups the data by splitting the ranges of both latitude and longitude into 20 equally spaced bins and creating factor variables Latitude_Binned and Longitude_Binned. For each combination of Department, year, month, Latitude_Binned and

Longitude_Binned the average Time.to.resolution and the total count is stored in variables Ave.Time.to.resolution and call_count.

Using this grouped data, your assistant then models the Ave.Time.to.resolution using two Poisson regression models, Poisson.1 and Poisson.2. The code for these models is provided.

(c) (3 *points*) Assess the differences between the two models, including fitted parameters, coefficient estimates, goodness of fit.

Candidates performed reasonably well on this task overall. Common reasons for partial credit were not correctly identifying the difference between the specifications of the two models and not comparing the goodness of fit of each model.

ANSWER:

Poisson.1 gives equal weight to each observation while Poisson.2 weights each observation by number of calls. This means observations with a small numbers of service calls contribute as much in fitting Poisson.1 as observations with a high number of calls.

Observations with a small number of calls generally have more variation since the target variable is averaged across fewer calls. Poisson.1 gives these observations relatively more weight than Poisson.2, leading to larger deviance. This is why Poisson.1 has a worse fit (AIC = 136.07) than Poisson.2 (AIC = 98.686).

Poisson.2 generally has coefficients that are closer to 0 than Poisson.1. Although both models find the intercept, DEPARTMENTSANITATION, and year to be significant, this leads to Poisson.1 having more significant p-values.