Exam PA April 18 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 9 tasks numbered 1 through 9 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. For this exam there is no data file, data dictionary, or .Rmd file provided. Neither R nor RStudio are available or required.

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.

A PDF is available labelled "Appendix." Some of the tasks/subtasks will reference this document. The reference will be in a red font and start with "See Appendix." The Appendix provides graphs, tables, or other output that will be needed for your answer.

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

Business Problem

You just started a rotation program with the New York City Economic Development Corporation and your first project is working with the team that manages New York City (NYC) ferries. The team is responsible for ferry schedules, ticketing and sales, and managing the ferry stops. You report to the Director of Data Analytics and are working with a junior analyst from the NYC Ferry team.

New York City was most impacted by COVID due to COVID-related shutdowns between March 2020 and December 2020. After the impact of the COVID-19 pandemic, the NYC Ferry team believes their previous models that were based on pre-pandemic data may no longer be valid. They are looking for you to help support rebuilding their models for the following purposes:

- *service demand planning*
- *employee hiring*
- *ticketing and sales support*
- *ferry stop vendor management.*

They are also interested in understanding the impact of the pandemic on ridership.

Your boss directs you to use a dataset[1] of public data that includes all ridership data from January 2019 – October 2022. There were about 320,000 service requests in this time period. Your assistant has prepared the public data and has provided the following data dictionary that contains all the variables appearing in the data.

[1] *Source: New York City Economic Development Corporation*

Data Dictionary

Variable	Data Type / Range	Description
Boardings	Numeric: 0 to 946	The number of riders who boarded a ferry in a particular hour
Route	String: With values of “RW” and “SV”	RW is Rockaway; SV is Soundview
Direction	String: With values of “SB” and “NB”	NB is northbound; SB is southbound
Stop	String: With the name of each stop	Name of ferry stop where boarding occurred
Daytype	String: with values of “Weekend” and “Weekday”	Whether the service was operating on a weekday or weekend/modified schedule.
Hour	Numeric: 0 to 23	Hour when riders are boarding in 24-hour clock. If value is 6, this refers to all boardings between 6:00 am-7:00 am. A value of 18 refers to boardings between 6:00 pm-7:00 pm.
Year	Numeric: 2019 to 2022	Year that boarding occurred
Month	Numeric: 1 to 12	Month that boarding occurred
Day	Numeric: 1 to 31	Day of month that boarding occurred

Task 1 (6 *points*)

To better optimize the staffing for the different stations, the NYC Ferry Team wants to understand how the number of boardings differs by stop.

Your assistant creates two graphs and wants to choose the graph that provides the more easily understood visualization of the relative number of boardings at each stop.

See Appendix – Task 1 Part a.

(a) (2 *points*) State which graph your assistant should use and explain why this graph is better than the alternative.

Candidates performed well on this subtask overall. Reasons for partial credit or no credit included only discussing the graphs and not selecting one, only discussing one graph, or incorrectly stating that the pie chart is better for comparing proportions.

ANSWER:

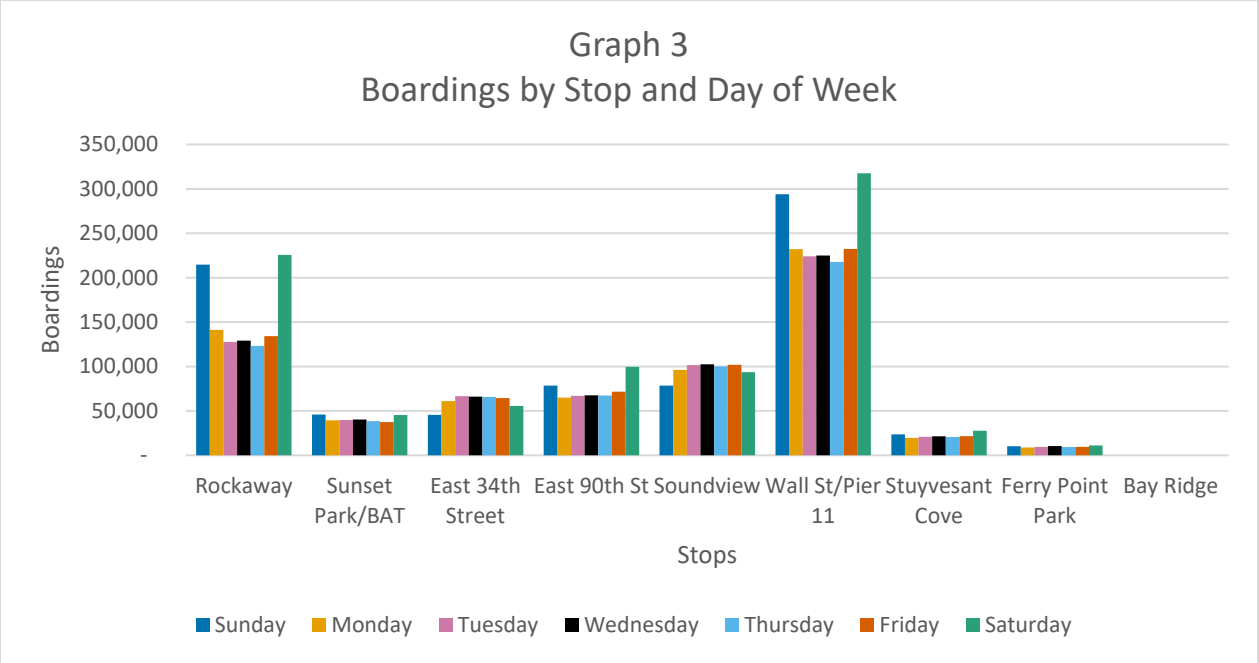
The assistant should choose Graph 1, the bar graph, rather than the pie chart.

Values in Graph 1 can be compared using length and position, whereas values in Graph 2 can be compared using angle or area. People are best at interpreting relative positions that are presented on a common scale, like Graph 1. By contrast, people are more prone to misjudge quantities that are encoded as angles or areas.

Additionally, the number of boardings at each stop can be read from graph 1 and the proportions can also be compared. The Number of boardings cannot be read from graph 2.

Your assistant trains a GLM where the target variable is the number of boardings. The model includes a calculated numerical field for **Daytype** (the values of 0 for weekdays and 1 for weekends) and this variable is statistically significant.

Your assistant wants to increase the granularity of the **Daytype** variable and replaces it with a numeric variable called **DayofWeek** that has values of 1 for Sunday, 2 for Monday, until 7 for Saturday. Your assistant finds that this variable is not statistically significant. Your assistant creates Graph 3 below to better understand what's going on.



(b) (*2 points*) Explain, using the graph above, why the **Daytype** variable is statistically significant while the **DayofWeek** variable is not.

*Candidate performance was mixed on this subtask, with most candidates receiving partial credit. Full credit responses discussed both how the non-linear relationship with **DayofWeek** leads to a statistically insignificant coefficient and why the same is not true for the **Daytype** variable.*

ANSWER:

There is not a monotonic, much less linear, relationship between ***DayofWeek*** and the number of boardings. A positive coefficient would create higher predictions on Saturday than Sunday, and a negative coefficient would have the opposite effect. However, we see from the chart that boardings are very similar for Saturday and Sunday. Therefore, the model will fit values close to zero, which are not statistically significant.

There is a clear relationship between ***Daytype*** and boardings, with ***Daytype*** = 0 having lower boardings overall than ***Daytype*** = 1. Although the relationship between ***Daytype*** and boardings varies by stop (for example, Soundview has lower boardings on the weekend than weekdays), there is a strong enough overall to fit a statistically significant coefficient.

(c) (2 points) Recommend two modeling enhancements that your assistant could explore based on Graph 3 above.

Candidate performance was mixed on this subtask, with most candidates receiving partial credit. In addition to the response in the model solution, recommending a tree-based model was a common full-credit response.

ANSWER:
My first recommendation is to change **DayofWeek** to a factor variable. This will allow the model to fit a different coefficient for each day of the week. This allows the model to fit to the non-linear relationship between **DayofWeek** and boardings at the expense of additional model complexity since the model needs to fit six coefficients instead of one.

My second recommendation is adding an interaction term between **DayType** and Stop.
This allows the model to reflect the fact that weekend boardings are higher than weekday boardings at some stops while the opposite relationship holds at other stops.

Task 2 (4 points)

In their initial data exploration, your assistant suggests applying Principal Components Analysis (PCA) as the main data exploration technique, noting the large number (about 320,000) of service requests present in the data.

(a) *(2 points)* Explain how PCA is typically used.

Candidate performance was mixed on this task. Many candidates only described the mechanics of PCA, with little or no discussion of how PCA is used. These responses were awarded partial credit. Full-credit responses described a way that PCA is used. In addition to the use for feature generation described in the model solution, credit was awarded for other uses of PCA, with data exploration being the most common alternative use being awarded credit.

ANSWER:

Principal component analysis is an unsupervised learning technique that creates new uncorrelated variables that maximize variance. Often, the first few principal components explain most of the variability in the original variables. These principal components can be used in place of the original variables to reduce dimensionality and create a simpler model.

(b) (*2 points*) Critique your assistant's suggested use of PCA with respect to the dataset.

Candidates struggled with this task overall. Many candidates described general weaknesses of PCA without any connection to the dataset or business problem. Most full-credit responses included a discussion similar to one or both of the complications described in the model solution.

ANSWER:

PCA and other unsupervised learning techniques often can be used in data exploration. However, it is not appropriate for this dataset:

- The dataset has small dimensionality. While there are many requests within the data, as our assistant notes, there are only 9 variables. PCA is effective when there is high dimensionality (many variables) which can make univariate and bivariate data exploration and visualization techniques less effective. PCA is used to summarize high-dimensional data into fewer composite variables while retaining as much information as possible. As we do not have high dimensionality, any information loss from feature transformation will not be outweighed by improvements in model performance or capture of latent variables.
- The dataset includes a significant number of factors variables, which will require conversion to numeric values prior to applying PCA. PCA attempts to maximize the variance or spread in our data distribution by linearly combining original variables.

Task 3 (11 *points*)

The NYC Ferry team is interested in building a model to predict boardings per day by ferry station. As a start, your assistant cleaned the data, split the data into training and testing sets, and created an ordinary least squares model.

(a) (*2 points*) List three assumptions for ordinary least squares regression with respect to residuals.

Most candidates received full credit on this subtask. In addition to the three assumptions listed below, credit was awarded for listing assumptions that residuals have no autocorrelation. Partial credit was awarded for vague, non-technical descriptions of the assumptions; a common candidate response receiving minimal partial credit was that the residuals show "no patterns." No credit was awarded for assumptions not related to residuals like the linearity relationship assumed between the predictors and predictions.

ANSWER:

Three assumptions of ordinary least squares regression are:

- The residuals have a normal distribution.
- The mean of the residual is zero.
- The residual variance is constant. This is also referred to as homoscedasticity.

You are provided with diagnostic plots from your assistant's OLS model.

See Appendix – Task 3 Part b.

(b) (*3 points*) Evaluate your assistant's model with respect to the three residual model assumptions from (a) based on the plots provided.

Generally, candidates who performed well on (a) also performed well on (b). Similar to (a), many candidates received partial credit for overly vague or non-technical responses. A common poor response was describing patterns observed in the plots without connecting explaining why the pattern is consistent with or validates one of the assumptions from (a).

ANSWER:

From the diagnostic plots, all three assumptions listed above are violated:

- Q-Q Plot shows that the distribution of residuals is not normal and has a fat right tail.
- The Residual vs. Fitted plot shows that the mean of the residuals deviates from 0 when the fitted value is large.
- Both the Residual vs. Fitted plot and Scale-Location plot show that the residual variance increases as the fitted value increases.

Your boss suggests to treat date variables (**Year**, **Month**, and **Day**) as categorical variables instead of numeric variables, and to test whether removing **Day** improves the model. Two additional models were built based on this suggestion:

- Model 1: **Year**, **Month**, and **Day** as categorical variables
- Model 2: Same as Model 1, but remove **Day** from the model

The code to create the two models and model outputs is provided.

See Appendix – Task 3 Part c.

(c) (*3 points*) Recommend either Model 1 or Model 2 to your client. Justify your recommendation.

Most candidates received full credit on this subtask. Although recommendations for either model were awarded credit provided sufficient justification, most candidates recommended Model 2 with justifications similar to the model solution. Candidates who did not receive full credit typically failed to recommend a model or recommended using another model entirely.

ANSWER:

I recommend Model 2, which removes the Day categorical variable. In reviewing the Model 1 output, the day categorical variable has 30 levels, but few of them are statistically significant. This suggests that there is not a strong relationship between Boardings and the Day variable. As expected, Model 1 has lower residual standard error and lower MSE on the training data. However, Model 2 has lower MSE on the testing data. This suggests that Model 1 is overfitting to a relationship between Day and Boardings that exists in the training data, but not in the testing data. Also, Model 1 is much more complex since it includes 30 additional categorical variables. The added complexity is not justified by the model's performance on the testing data.

Your client is interested in forecasting the daily boardings in 2023. Your assistant tried to run the predictions using Model 1 and Model 2, but both models ran into errors.

(d) (*1 point*) Explain why neither Model 1 nor Model 2 can make predictions for dates in 2023.

Most candidates received full credit on this subtask. Full credit was awarded for identifying both that Year is a categorical (or factor) variable, and that the level 2023 does not exist in the training data.

ANSWER:

Year is modeled as a categorical variable in both models. The training data has only 4 levels for this variable: 2019, 2020, 2021 and 2022. Since the trained model did not estimate a coefficient the 2023 level, it cannot make a prediction for dates in 2023.

You are provided with various plots.

See Appendix – Task 3 Part e.

(e) (2 points) Recommend three modeling improvements with reference to the model output and plots.

Candidate performance was mixed on this subtask, with most candidates receiving partial credit. Most candidates provided a recommendation based on the skewness of the target variable, and many candidates also made a recommendation around seasonality. Any clear and accurate recommendation based on the output provided was awarded credit. No credit was awarded for duplicate recommendations that paraphrased a prior recommendation.

ANSWER:

I recommend the following improvements based on the model output:

- The distribution of **boarding.day** is right-skewed. I recommend applying a log transformation to the **boarding.day** variable or using a GLM with a right-skewed distribution like Poisson or gamma.
- The distribution of **boarding.day** by Year shows a nonlinear relationship between Year and **boarding.day,** with **boarding.day** lowest in 2020. This may be due to COVID-related shutdowns. Therefore, I recommend removing 2020 data.
- The distribution of **boarding.day** by Date shows seasonality within each year. I recommend adding a seasonality variable to the model.

Task 4 (4 *points*)

Your boss asks you to create a generalized linear model to help determine which ferry stops have a meaningful impact on the reduction in boardings in 2021 compared to 2019. You produce the following output:

Initial GLM Summary:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.549509   0.039451 -13.929  < 2e-16 ***
weekend_ind              0.174105   0.019023   9.152  < 2e-16 ***
Month                    0.025141   0.002541   9.894  < 2e-16 ***
northbound_ind           0.114409   0.022354   5.118 3.32e-07 ***
rockaway_ind            -0.104541   0.045118  -2.317   0.0206 *
Stop.East.34th.Street -0.053341   0.035684  -1.495   0.1351
Stop.East.90th.St        0.327681   0.035634   9.196  < 2e-16 ***
Stop.Rockaway           -0.115437   0.045019  -2.564   0.0104 *
Stop.Soundview           0.030479   0.045396   0.671   0.5020
Stop.Sunset.Park.BAT     0.182429   0.035338   5.162 2.63e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You ask your assistant to reduce the number of variables in the model using an algorithmic feature selection method and they propose using ridge regression and not lasso regression.

(a) (*2 point*) Critique your assistant's proposal.

Candidates performed well on this subtask. Full-credit responses recognized that ridge regression cannot entirely remove features, contrasting it with lasso or elastic net with alpha greater than zero.

ANSWER:

The assistant's proposal to use ridge regression would not result in removing any variables from the model. While ridge regression does use a penalty to reduce the size of coefficients, its penalty function cannot reduce coefficients to zero, meaning it cannot remove variables from the model. However, lasso regression uses an absolute value penalty and can reduce a coefficient to zero, removing it from the model and performing feature selection.

Your assistant performs regularized regression and changes the mixing coefficient (alpha) parameter. Each model uses the optimal lambda value for that value of the mixing coefficient. The model coefficients are shown below:

	Model 1	Model 2	Model 3
(Intercept)	0.676967945	0.7290097575	0.76214890
weekend_ind	0.120324669	0.0758948800	0.02550108
month_transform	0.013206528	0.0003158021	.
northbound_ind	0.073389666	0.0236195285	.
rockaway_ind	-0.072772576	.	.
Stop.East.34th.Street	-0.054687152	.	.
Stop.East.90th.St	0.241379707	0.2287893882	0.18490437
Stop.Rockaway	-0.089193413	-0.0450787536	.
Stop.Soundview	-0.005870593	.	.
Stop.Sunset.Park.BAT	0.093633905	0.0045934356	.

(b) (2 *points*) Complete the chart below with the name of each type of regularized regression model, the possible value or values of the mixing coefficient (alpha) that could produce each model, and one benefit of each type of regularized regression model.

Candidate performance was mixed on this subtask, with most candidates receiving some amount of partial credit. Full-credit response indicated that Model 1 has the lowest alpha and Model 3 has the highest alpha or provided plausible values for alpha. Many candidates struggled to explain the benefits of different values of alpha.

ANSWER:

	Model 1	**Model 2**	**Model 3**
Type of Regularized Regression:	Ridge or Elastic Net Regression	Elastic Net Regression	Lasso or Elastic Net Regression
Mixing Coefficient (alpha) Value(s):	$0 \leq$ alpha < 1 and alpha is smaller than for Model 2	$0 <$ alpha < 1 and alpha is larger than for Model 1	$0 <$ alpha ≤ 1 and alpha is larger than for Model 2
Benefit:	Reduces variance by shrinking coefficients.	Reduces variance by shrinking coefficients, can also be used to perform model selection and is helpful in instances where there is high-dimensional data with few data points.	Reduces variance by shrinking coefficients and can also be used to perform model selection and remove nonpredictive variables.

Task 5 (9 *points*)

Your manager is interested in understanding the impact that the COVID-19 pandemic had on ferry ridership in 2020. You are asked to build a decision tree to address this question.

(a) (3 *points*) Compare and contrast single decision tree and tree-based ensemble models.

Candidates performed well on this subtask. Most candidates effectively contrasted the models. Some candidates failed to compare the models, based their response on characteristics of a particular tree-based ensemble model rather than those models in general, or provided incorrect explanations of the bias/variance tradeoff.

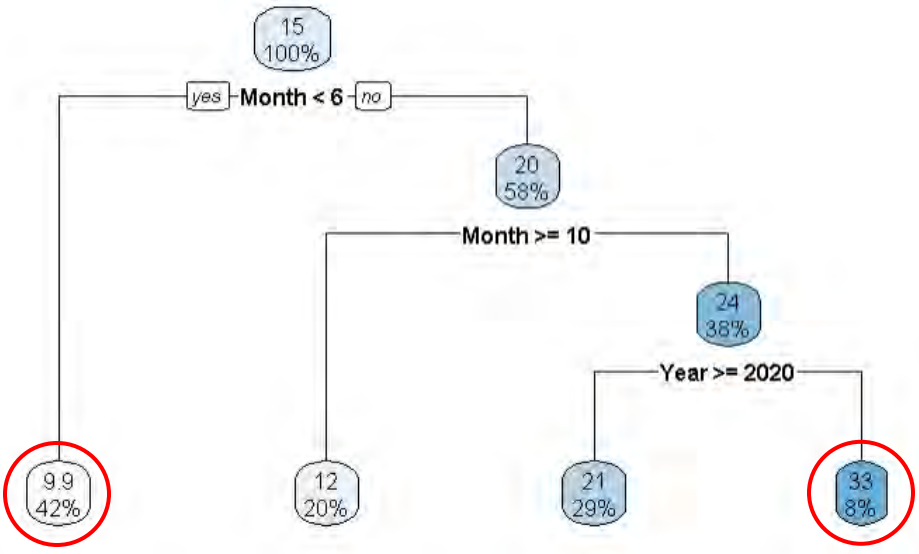
ANSWER:

Both methods can be used to build regression or classification models, determining splits based on impurity or information gain measures.

A single decision tree partitions the data into a single set of non-overlapping regions. Ensemble methods (such as random forests and boosted trees) fit multiple decision trees and combine the predictions from the trees fit to determine the model's prediction.

Singular trees are prone to high variance due to overfitting, while ensemble methods reduce overfitting.

You create a decision tree to predict hourly ferry boardings. The resulting tree is printed below.



(b) (2 *points*) Interpret the left and right terminal nodes in the model.

Most candidates performed well on this subtask, providing correct interpretations of both nodes. Candidates were expected to explain how to reach each node and interpret the values at that node. The most common errors were with interpretation of the rightmost node.

ANSWER:

Interpretation of the left terminal node: In months 1-5, there are an average of 9.9 ferry boardings each hour. This node accounts for 42% of total records in the training data set.

Interpretation of the right terminal node: In months 6-9, before 2020, there were an average of 33 ferry boardings each hour. This node accounts for 8% of total records in the training data set.

(c) (*2 points*) Determine if this tree shows an interaction between month and year. If there is an interaction, describe it. If not, explain why there is no interaction.

Candidate performance was mixed on this subtask, with various reasons for candidates receiving partial or no credit. No credit was awarded to candidates who stated without explanation that there was no interaction or stated that trees cannot capture interactions. Many candidates correctly identified the node where interaction occurred in this tree but were unable to describe the resulting interaction.

ANSWER:

The node representing months 6-9 has a subsequent split around 2020, which did not occur for other months. This indicates an interaction between the year and month variables. During months 6-9, the average daily ferry boardings in 2019 were 33. In 2020 and after, the average daily ferry boardings were 21. It seems like there was a significant decrease in average ferry boardings before and after the onset of COVID-19 for the summer months.

After reviewing your decision tree, your manager concludes that COVID-19 did not have a material impact on average hourly ferry boardings.

(d) (*2 points*) Explain why your manager may have come to this conclusion. State whether you agree or disagree with this conclusion and justify your choice.

Candidate performance was mixed on this subtask, with most candidates receiving partial credit. No credit was given if a candidate didn't provide a reason their manager may have come to that conclusion. Some candidates questioned the data and the modeling methodology; these discussions generally were not awarded any credit. Partial credit was awarded for valid statements about COVID-19 that lacked reference back to the tree.

ANSWER:

My manager may have come to this conclusion given that the year variable is used as a split far down the tree. This suggests that there is only a noticeable differentiation in average ferry boardings for a small subset of months.

I disagree with my manager's statement. Although the year split happens further down the tree than month, that does not indicate that it is not material. In isolation, the difference in mean ferry boardings between months (e.g., from seasonality) is more significant than year-over-year differences. This illustrates one of the limitations of decision trees: They make greedy splits based on the largest information gain at each immediate step, not necessarily to produce the best fitting overall model. Further, COVID-19 impacts would have both a year and month component. Therefore, further analysis is required to determine if there is an impact from COVID-19.

Task 6 (4 points)

To optimize staffing of shifts for the different stops, you want to understand how the boardings are distributed throughout the day.

Your assistant is unsure whether to use the **Hour** variable directly from the dataset or whether to use a newly created variable, **TimeofDay**, which is based on a less granular grouping of hours. The categorical variable **TimeofDay** takes the values of "morning" (hours 5-10), "afternoon" (hours 11-16) and "evening" (hours 17-22).

Looking at one group of stops, your assistant creates a line graph showing each stop's proportion of riders at each **Hour**, as well as a bar graph showing each stop's proportion of riders at each value of **TimeofDay**.

See Appendix – Task 6 Part a.

(a) (2 *points*) Describe a strength of each graph relative to the other.

Candidate performance was mixed on this subtask. Credit was awarded for any valid strength in the business context provided.

ANSWER:

Graph 1 contains more granular information than Graph 2, allowing us to see how boardings vary both within and between the less granular times of day. This would be more useful in determining precisely when shifts should begin and end. For example, in the morning, the graph shows us that there are more boardings at 7 than at 5 and 6.

Although Graph 2 contains less granular information, it also has less noise, making it easier to interpret broad trends. For example, we can see that Soundview has a much higher proportion of boardings in the morning hours than the other stops, meaning that shifts at Soundview should be scheduled differently from the other stops.

When doing exploratory analysis on the other stops, your assistant notices that the graphs look very different depending on whether the graph is based on the proportion of boardings or if it's based on the number of boardings.

See Appendix – Task 6 Part b.

(b) (2 points) Describe a strength of each graph relative to the other.

Candidates performed better on this subtask than on (a). Credit was awarded for any valid strength in the business context provided.

ANSWER:

Graph 1 allows us to clearly see how ridership varies across hours within each stop. This is helpful if we have already determined the number of staff at each stop and need to decide when to schedule their shifts.

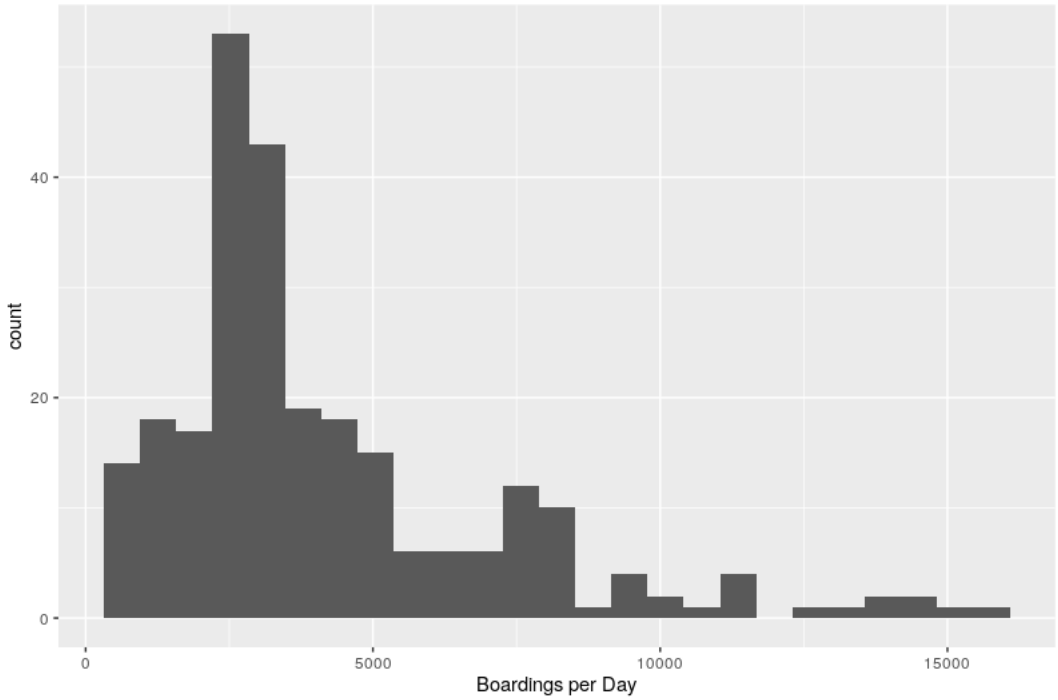
Graph 2 has the advantage of allowing one to interpret total ridership between stops, making it clear that Wall St/Pier 11 is much more widely used than the other stops. However, since the graph is scaled based on the data including Wall St/Pier 11, it is much harder to see the distribution of boardings of the other stops. This is more useful if staff can be easily reassigned from one stop to another.

Task 7 (6 points)

Your boss has asked for your support in estimating how many fewer passengers are using the ferry as of October 2022 compared to what would have been expected based on 2019 ridership, were the COVID pandemic not to have occurred.

You decide to construct a generalized linear model (GLM) by limiting the dataset to calendar year 2019 and randomly assigning 70% of the data to a training set and 30% to a testing set. Then, you produce the following graph with the distribution of the number of daily boardings:

Graph is displaying the distribution of Boardings per Day in the training set



(a) (*2 points*) Identify a model distribution that would be a reasonable choice and one that would not be a reasonable choice for this data. Justify your choices.

Candidates did well on this subtask overall, with most candidates receiving full credit. Full-credit responses clearly identify a reasonable and unreasonable distribution along with sufficient justification as to why those distributions were chosen. Some candidates identified link functions and not distributions. These responses were not awarded credit.

ANSWER:

Gamma would be a reasonable choice of model distribution because the variable has only positive numbers and has right skew.

The binomial distribution would not be a reasonable choice since the variable is not binary or categorical.

Your assistant decides to create two ordinary least squares models with one variable each: the first uses month number as a numeric variable and the second creates a new feature that takes the absolute value of the difference between the month and 7.5: | month – 7.5 |. The graphs provided show observed boardings per day (purple) vs. predicted boardings per day (green) for each model.

See Appendix – Task 7 Part b.

(b) (2 *points*) Explain how the variable transformation affected the predicted boardings per day in each model.

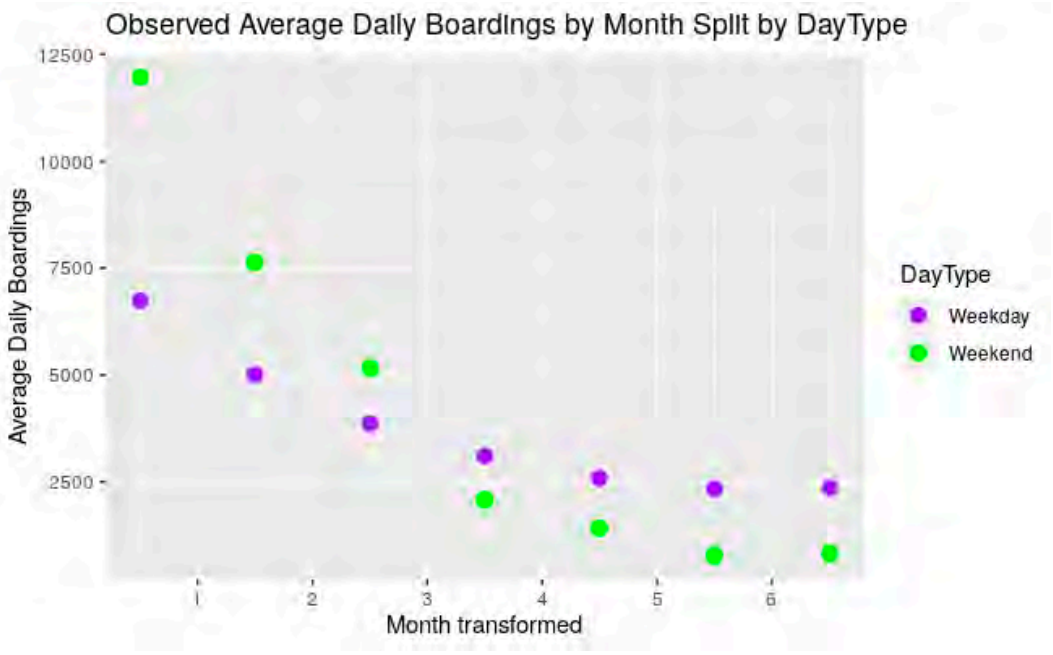
Most candidates received partial credit on this subtask. A significant number of candidates failed to discuss both models. Many candidates did not explain how the transformation resulted in the increasing, then decreasing shape of the predictions.

ANSWER:

In the first model the month variable is an untransformed numeric variable. The model therefore needs to fit a linear, monotonic relationship, which clearly doesn't fit the distribution of Boardings by Month. The predicted boardings per day show a slight upward trend in the linear predicted line.

In the second model the month variable Is modeled as the absolute value of the difference between month and 7.5. This enforces a stepwise, V-shape on the predicted boardings per month. The boarding data leads to a downward-opening V and sets the overall level and the size of the steps. This variable transformation fits the nonlinearity of the data better, resulting in more accurate predictions.

You decide to move forward with creating a GLM using a distribution that is appropriate for the data. Given the moderate linear relationship between daily boardings and the generated month feature, you decide to start with that variable. You hypothesize that **DayType** may also be important and ask your assistant to create a GLM with that variable. They inquire about whether they should also include an interaction variable for Month transformed and **DayType**.



(c) (*1 point*) Explain what an interaction variable might capture in this case that wouldn't otherwise be caught by the model.

Candidate performance was mixed for this question. Strong answers addressed the relationship and established the kind of interaction that it would capture, while clearly explaining how the interaction would be different from the other variables alone. Many candidates simply defined an interaction, without any connection to the information provided.

ANSWER:

We see that the slope of average daily boardings is higher for Weekend than Weekday. An interaction would allow the model to estimate different transformed month coefficients for each DayType. Without an interaction, the model would use the same transformed month coefficient for both levels of DayType.

After reviewing your models, your assistant hypothesizes that day of the week may also be important. They suggest adding in indicator variables for specific days of the week. Your assistant adds the variables and creates a model with the following output:

```
Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               9.083410   0.171276  53.034  < 2e-16 ***
weekend_ind               0.641827   0.165308   3.883 0.000133 ***
Month                    -0.024517   0.007227  -3.392 0.000807 ***
month_transform          -0.213901   0.015508 -13.793  < 2e-16 ***
month_t_day_interaction -0.310638   0.026351 -11.788  < 2e-16 ***
monday_ind               -0.014546   0.163231  -0.089 0.929067
tuesday_ind              -0.089435   0.170420  -0.525 0.600199
wednesday_ind            -0.073245   0.165772  -0.442 0.658992
thursday_ind             -0.082102   0.163301  -0.503 0.615579
friday_ind               -0.005076   0.166066  -0.031 0.975642
saturday_ind              0.282645   0.084539   3.343 0.000957 ***
sunday_ind                      NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(d) (*1 point*) Interpret the NA values in the **sunday_ind** variable.

Results were mixed for this question. Full credit was given if the candidate identified that the reason was due to multicollinearity, identified that Sunday was a baseline, or noted that the weekend_ind variable introduced collinearity. The most common response receiving no credit was stating that the ferries must not run on Sundays. Note that with weekend_ind and all seven day of week indicator variables, two variables should have produced NA values; however, an issue with the model resulted in only one NA being produced.

ANSWER:

Each day of the week coefficient represents the difference in the linear prediction relative to a base level of the day of the week variable. By including a variable for every day of the week in the model, we have introduced perfect multicollinearity. The NA values exist because only 6 day of week coefficients are required for a categorical variable with 7 levels.

Task 8 (15 points)

(a) (*3 points*) Compare and contrast stepwise selection with shrinkage methods such as lasso and ridge.

Candidate performance was mixed on this subtask with most candidates receiving partial credit. Candidates were required to both compare and contrast the methods for full credit. The most common reasons for partial credit were (1) only contrasting or comparing the methods, (2) comparing and contrasting ridge and lasso, and (3) comparing and contrasting forward and backward selection.

ANSWER:

Similarities

- Both stepwise selection and lasso/ridge regression avoid overfitting to the data, especially when the number of observations is small compared to the number of predictors.
- Both stepwise selection and lasso regression can be used for variable selection to reduce model complexity.

Differences

- Stepwise selection takes iterative steps, either from no predictors (forward selection), or from a model with all predictors (backward selection). The selection process adds or remove predictors until there is no improvement as measured by AIC, or another performance metric.
- Shrinkage methods fit coefficients for all predictors simultaneously to optimize a loss function that includes a penalty parameter that penalizes large coefficients. Shrinkage methods can reduce the size of coefficients without entirely eliminating variables, which stepwise selection cannot do.

(b) (*1 point*) Explain why variables are standardized as part of the lasso model fitting procedure.

Candidate performance was mixed on this subtask with most candidates receiving partial credit. Many candidates only partially explained standardization's impact on model fitting or used overly-vague language.

ANSWER:

The lasso regulation term is the sum of absolute values of model coefficients. Variables that are on a larger scale typically have smaller coefficients and vice-versa. Without standardizing, the regularization will focus on shrinking the variables on a smaller scale over those on a larger scale.

(c) (*2 points*) Describe the process of searching for the optimal value of the hyperparameter lambda in a lasso regression.

Candidates performed well on this task overall. Full credit was awarded for a complete description of the cross-validation process together with an explanation of how it is applied in the context of selecting lambda for a lasso model. Partial credit was awarded for descriptions of cross-validation without discussing how it is used for selecting lambda.

ANSWER:

The optimal value for lambda can be found using cross-validation. First, a grid of lambda values is chosen for the search. Then for each lambda value, a cross-validation error is calculated.

The first step in calculating a cross-validation error is to partition the data into k folds. A single fold is removed for testing, and the remaining folds are used to train a lasso model with the current lambda value. This process is repeated for each of the k partitions, and a cross-validation error is calculated as the average of an error measure (e.g. RMSE or AUC) across all k testing partitions.

The optimal lambda value is the one with the lowest cross-validation error.

(d) (*1 point*) Describe how the lambda hyperparameter impacts variable coefficients in a lasso regression.

Candidates performed well on this task overall. Full credit responses clearly described the mechanics of how lower and higher lambda values impact model coefficients.

ANSWER:

A larger lambda increases the weight given to the size of each coefficient when calculating the penalty term. This forces the model to "shrink" the coefficient values more than for smaller lambda values. Large enough values of lambda will shrink all coefficient values to zero, resulting in a model that consists of only an intercept.

To better plan ticket office staffing at each ferry stop, your client wants to predict whether the number of boardings in the next hour will be greater than 150 based on past hour boarding data from all stops. Your assistant helped to clean and prepare the data for modeling.

You are provided with exploratory analysis of the new variable and model output from 3 different models.

- Model 1: GLM with all possible variables including **Month**, **Day**, **Hour**, **Daytype**, **Stop** and **Boardings** at each of 8 stops in the past hour. **Month**, **Day** and **Hour** are modeled as categorical.
- Model 2: A backward selection run on Model 1.
- Model 3: Lasso with the same set of variables as Model 1. Lambda is set to be 0.0004.

See Appendix – Task 8. Part e.

(e) (*4 points*) Compare the model results and recommend a model to your client. Justify your recommendation.

Candidates performed well for this task. Full credit responses provided a comparison based on the model output provided and also made a clear recommendation that considered both model performance and the business problem. Some candidates failed to link the business problem with their recommendations and only chose one model without right supporting statements.

ANSWER:

Model Comparison:

- Model 1 is a logistic regression with all variables. This model has the highest AUC in the training set. The test set AUC is relatively close to training AUC, meaning the overfitting is marginal. However, from the model coefficient table, most of the Day levels and Stop are not statistically significant.

- Model 2 uses backward selection starting from Model 1. The stepwise selection drops the variable Day and two of the previous-hour boarding stops. This results in a slight improvement of AIC due to the reduction of model complexity. From an AUC perspective, Model 2 has a slight AUC drop on training data, but a gain in the test data.

- Model 3 uses lasso regression, adding a regulation term on model coefficients. With lambda = 0.0006, the model drops some of the Day levels and the same two previous-hour boarding stops variables compared to Model 2. The coefficients are also overall closer to zero than Models 1 and 2. The resulting AUC is lower than Model 1 and 2 on both training and test data.

Recommendation:

I recommend Model 2 to my client based on the following reasons:

- Model 2 has the highest AUC on the test data among all 3 models

- Model 2 has the least number of coefficients in the model, making it favorable in the implementation. For example, the client doesn't need to collect hourly boarding data for two of the eight stops.

You are provided with the confusion matrix produced by the lasso model with a positive response cutoff threshold of 0.5.

		Reference	
		Negative	Positive
Prediction	Negative	38,128	1,064
	Positive	148	728

(f) (*2 points*) Calculate sensitivity and specificity. Show all work.

Candidates performed well on this task, with most candidates receiving full credit. Partial credit was awarded for only correctly calculating one of the metrics, not showing calculation steps, or correct formulas with calculation errors.

ANSWER:

Sensitivity = $\frac{TP}{TP+FN} = \frac{728}{728+1064} = 0.406$

Specificity = $\frac{TN}{TN+FP} = \frac{38128}{38128+148} = 0.996$

Your boss recommends lowering the cutoff threshold.

(g) (*2 points*) Assess the consequences of this recommendation as it relates to the business problem.

Candidate performance was mixed on this task, with most candidates receiving partial credit. Most candidates were able to provide a technical explanation of the effects of changing the threshold, but very few were able to make a strong connection to the business problem.

ANSWER:

This will increase positive predictions (both TP and FP) while reducing negative predictions (both TN and FN), increasing sensitivity. In the context of the business problem increasing sensitivity is important as it gives the ticket office an indication that there is likely to be a large rush of customers, and they can respond by increasing staffing. Having higher sensitivity increases the likelihood that the ticket office will adequately staff when there is likely to be a lot of customers. However, lowering the cutoff threshold also reduces specificity, meaning there may be more instances when the ticket office is overstaffed and there are not a large group of customers.

Task 9 (11 points)

You are working with your manager to predict the number of Boardings and determine the key drivers of service demand. Your manager recommends starting with a random forest model.

(a) (3 *points*) Describe how bagging is used in the random forest algorithm and the advantage it gives random forests over a single decision tree in terms of the bias/variance trade-off.

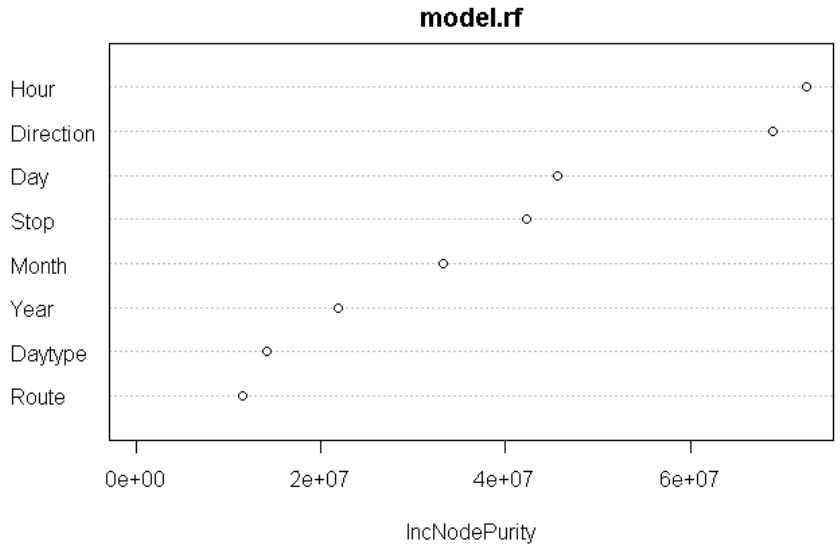
Candidate performance was mixed on this subtask, with most candidates receiving partial credit. Many candidates discussed ensemble methods in general rather than discussing bagging in a random forest.

ANSWER:

Random forests are created by applying bagging and taking random feature subsets to construct multiple trees, which are averaged to produce a prediction.

Bagging is the process of training of multiple models in parallel on different random subsets of the data. Each individual tree is trained on a different training dataset. Variance refers to the sensitivity of the model to changes in the training dataset. Bagging reduces variance because each individual tree is trained on different data..

Your manager is interested in determining the key drivers of Boardings. Your assistant built a random forest to model the number of Boardings and shares the following plot to help interpret the results.



(b) (*2 points*) Interpret the plot.

Candidates performed well on this subtask. Full-credit responses explained how to read the variable importance plot and correctly identified which variables are the most or least significant.

ANSWER:

This is a variable importance plot, which is one way of showing how much impact a variable has on model predictions. Variables with higher incremental node impurity have higher importance. The plot indicates that hour and direction are the two most significant drivers of Boarding count in this random forest model.

(c) (2 *points*) Describe how values for a partial dependence plot are calculated for a specific variable in a random forest model.

Candidates struggled with this subtask. Many candidates discussed the variable importance plot for (b) instead of partial dependence plots; these responses were not awarded any credit. Candidates who described parts of the calculation correctly received partial credit. The most common partial credit responses described some type of calculation where the value of the predictor variable is fixed at the average of that variable across all observations.

ANSWER:

To calculate partial dependence of a variable, we need to first replace all values for that variable in the dataset with the lowest value for that variable, calculate predictions for all observations, and average the predictions. This process is repeated for all values of the variable (or a selected grid of values). The partial dependence plot shows the predictor variable on one axis and the average predictions on the other axis.

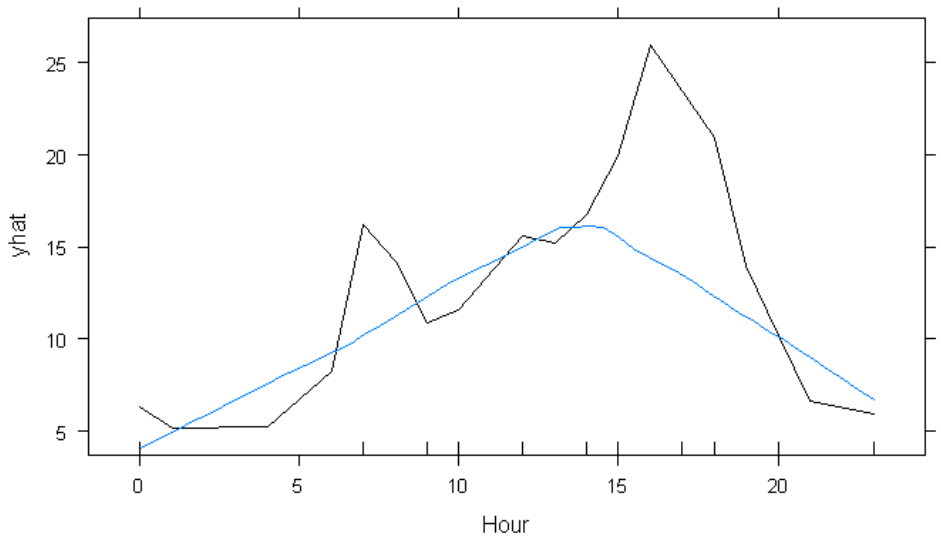
(d) (*1 point*) Identify one limitation of using a partial dependence plot to interpret the model.

Candidate performance was mixed on this subtask. Limitations other than the one in the model solution were awarded credit. Many candidates stated only that partial dependence calculations are computationally intensive; these responses were not awarded any credit.

ANSWER:

If two predictors are correlated, the PDP will calculate predicted values for unrealistic combinations while the model itself was only fit to realistic combinations. For example, if height and weight are predictor variables, making everyone seven feet tall will force predictions for individuals that tall yet weighing only 100 pounds.

Your manager asks you to provide additional detail about the relationship between **Hour** and **Boardings**. Your assistant creates the following partial dependence plot.



Your assistant added the blue smoothed line to the partial dependence plot, saying that it makes it easier to interpret the relationship between **Hour** and **Boardings**.

(e) (3 *points*) Recommend whether or not to include the smoothed line in the report to your manager. Justify your recommendation.

Candidate performance was mixed on this subtask. No credit was awarded for critiques of the chart itself without discussing the lines at all. No credit was awarded for dismissing the bimodality as a result of data quality issues or outliers. Partial credit was awarded to candidates who accurately identified disadvantages of including the smoothed line without making a recommendation.

ANSWER:

I do not recommend including the smoothed line in the report. The unsmoothed plot shows ferry boardings by hour have a bi-modal distribution, with one mode around hour 7 and another around hour 17. This might be important information to have when deciding when staffing needs are higher. By contrast, the smoothed plot results in a uni-modal distribution, which is misleading and may lead to poor staffing decisions.