Exam PA April 18 Project Statement

IMPORTANT NOTICE – THIS IS THE APRIL 18, 2023 PROJECT STATEMENT. IF TODAY IS NOT APRIL 18, 2023, SEE YOUR TEST CENTER ADMINISTRATOR IMMEDIATELY.

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 need 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 occured
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.

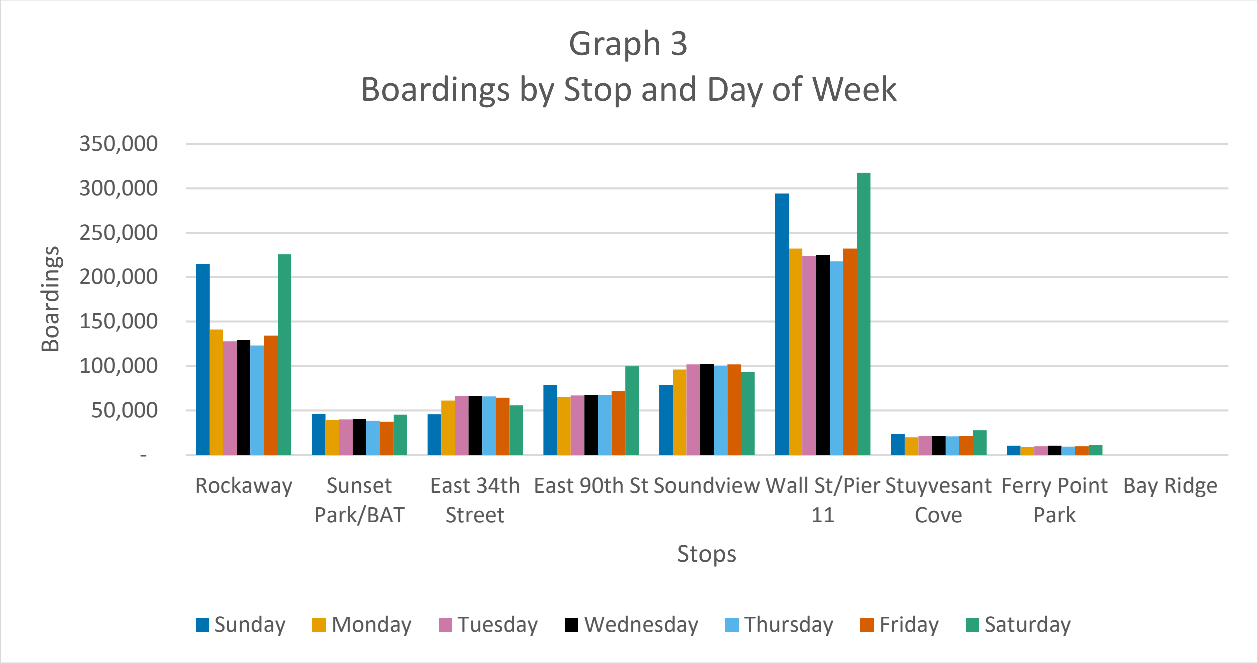
See Appendix – Task 1 Part a.

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

ANSWER:

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.

ANSWER:

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

ANSWER:

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.

ANSWER:

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

ANSWER:

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.

ANSWER:

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.

ANSWER:

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.

ANSWER:

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.

ANSWER:

You are provided with various plots.

See Appendix – Task 3 Part e.

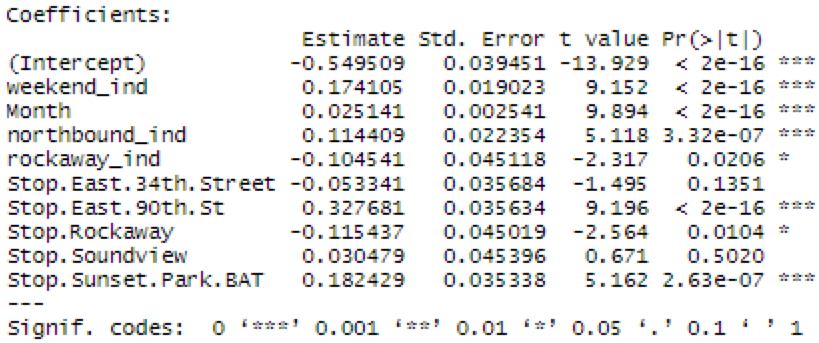
(e) (*2 points*) Recommend three modeling improvements with reference to the model output and plots.

ANSWER:

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:



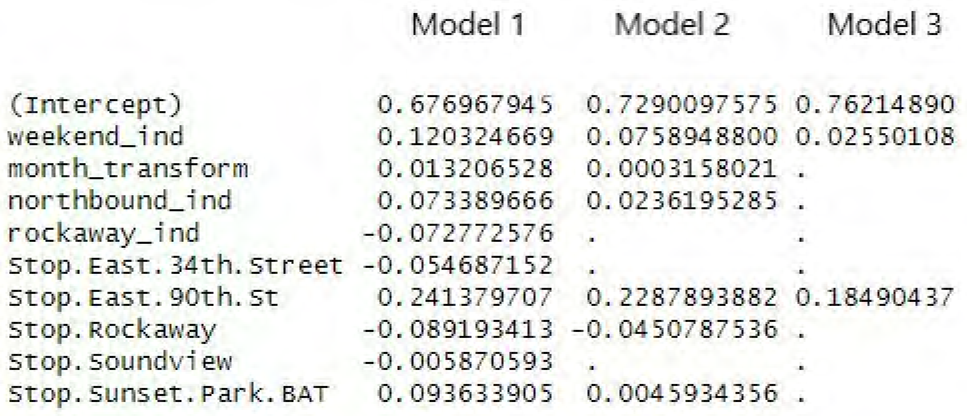
```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -0.549509   0.039451 -13.929  < 2e-16 ***
weekend_ind           0.174105   0.019023   9.152  < 2e-16 ***
Month                 0.025141   0.002541   9.894  < 2e-16 ***
northbound_ind        0.114409   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.

ANSWER:

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.

ANSWER:

	Model 1	Model 2	Model 3
Type of Regularized Regression:			
Mixing Coefficient (Alpha) Value(s):			
Benefit:			

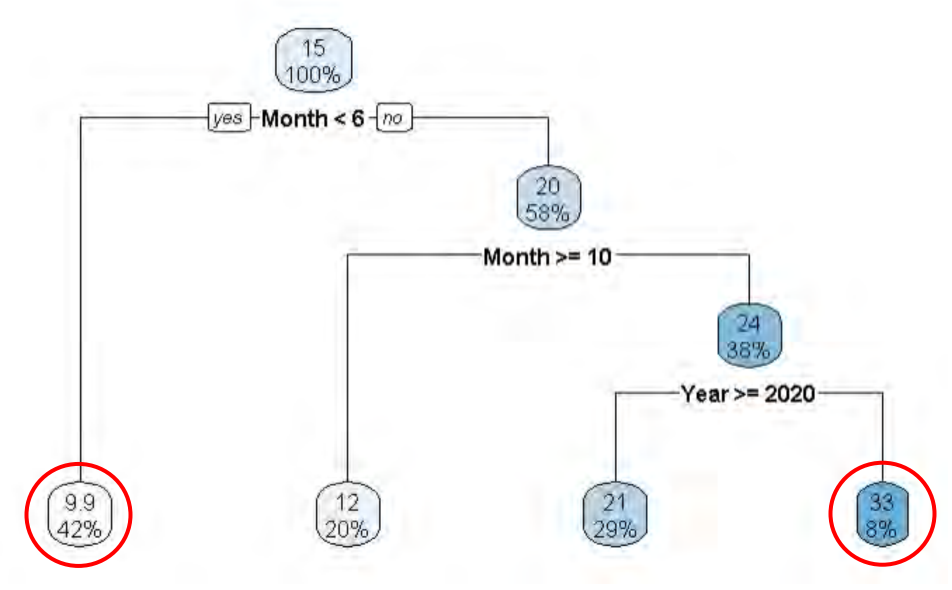
Task 5 (9 *points*)

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.

ANSWER:

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.

ANSWER:

(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.

ANSWER:

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.

ANSWER:

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.

ANSWER:

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.

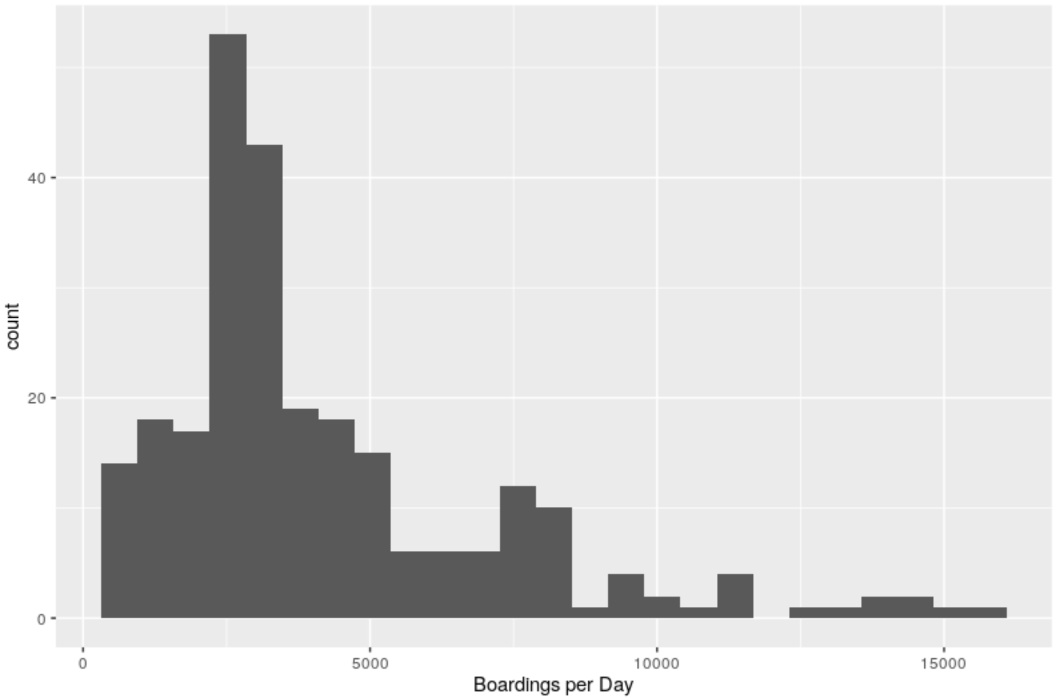
ANSWER:

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.

ANSWER:

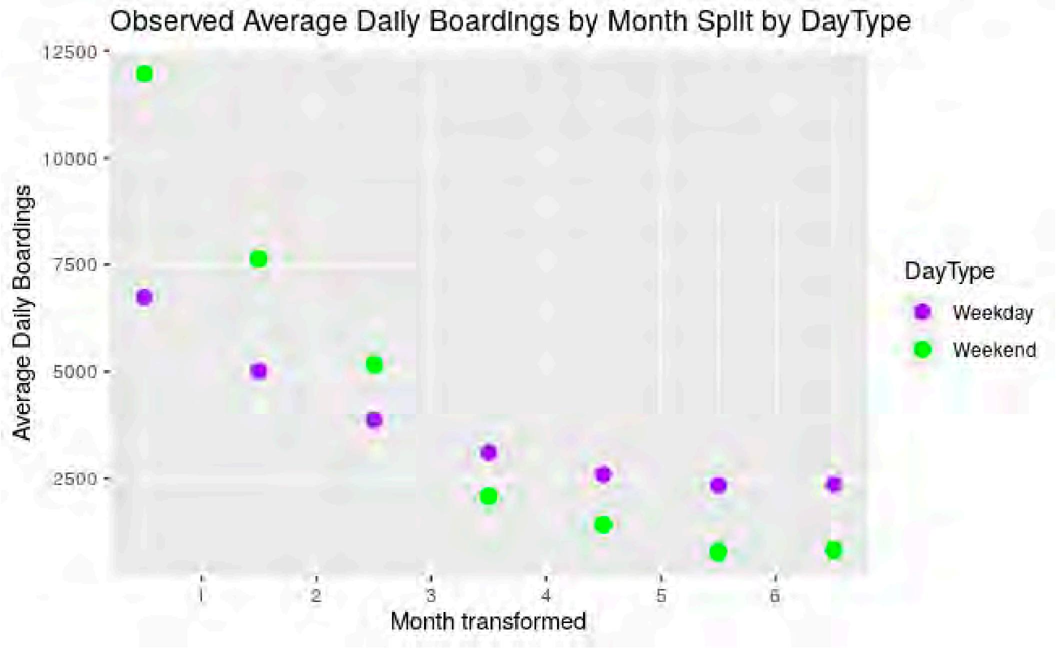
Your assistant decides to create two ordinary least squares models with one variable each: the first uses month number as a numeric variable and the second creates a new feature that takes the absolute value of the difference between the month and 7.5: | month – 7.5 |. The graphs provided show observed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

ANSWER:

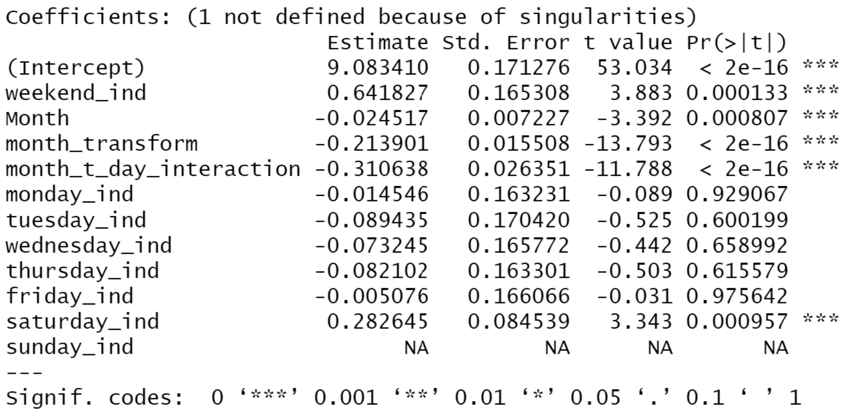
You decide to move forward with creating a GLM using a distribution that is appropriate for the data. Given the moderate linear relationship between daily boardings and the generated month feature, you decide to start with that variable. You hypothesize that **DayType** may also be important and ask 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.

ANSWER:

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.

ANSWER:

Task 8 (15 points)

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

ANSWER:

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

ANSWER:

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

ANSWER:

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

ANSWER:

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.

ANSWER:

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.

ANSWER:

Your boss recommends lowering the cutoff threshold.

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

ANSWER:

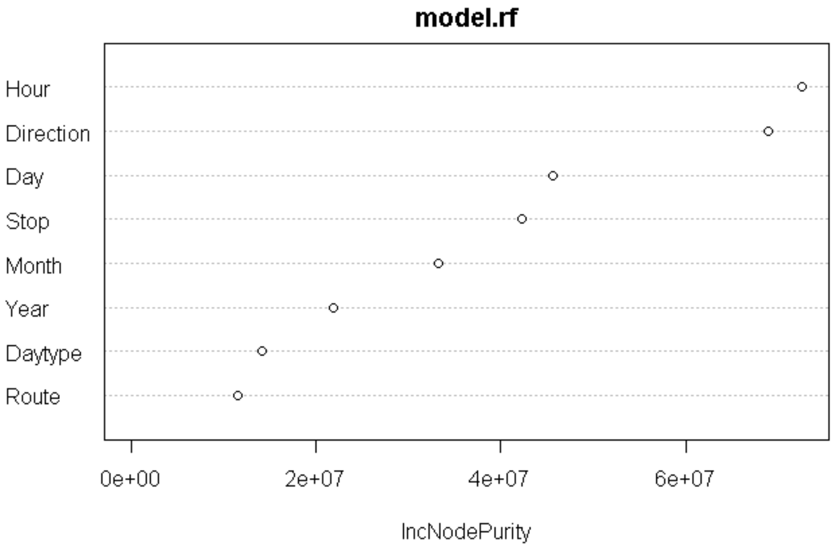
Task 9 (11 points)

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.

ANSWER:

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.

ANSWER:

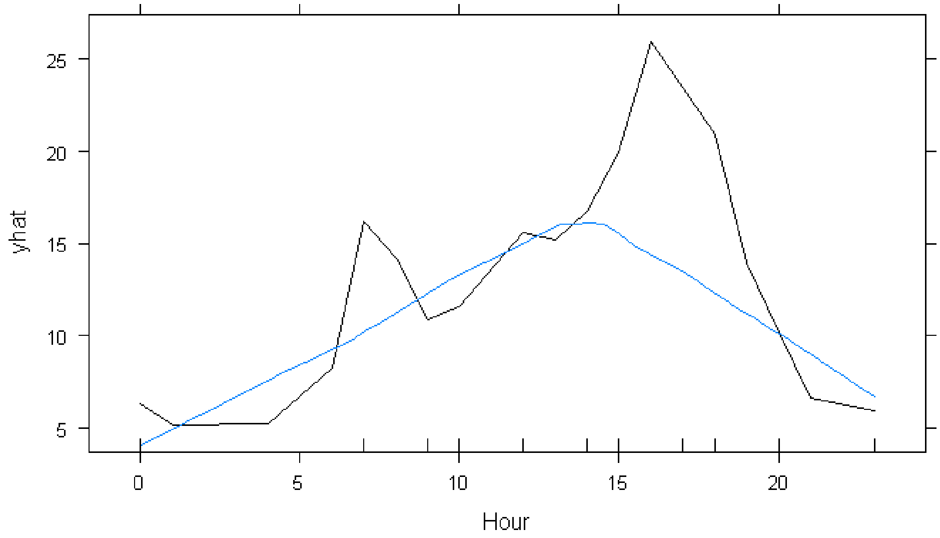
(c) (2 *points*) Describe how values for a partial dependence plot are calculated for a specific variable in a random forest model.

ANSWER:

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

ANSWER:

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.

ANSWER: