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.

June 21 2021 Exam PA - Model Solution

Instructions to Candidates: Please remember to avoid using your own name within this document or when naming your file. There is no limit on page count.

Also be sure all the documents you are working on have June 21 attached.

As indicated in the instructions, work on each task should be presented in the designated section for that task.

Task 1 – Define the business problem (5 points)

Quality responses to this question demonstrated knowledge of business problem definition including a description of the problem, a brief description of the data and its source and how a model will be evaluated.

Candidates performed well on this task. Well-prepared candidates were able to describe the data and note data elements were not sourced from the company. A minority of candidates wrote significantly more than was necessary for full credit.

Candidates failed to earn points if they did not mention the target and how the travel agency would incorporate the model into their business process.

A Canadian travel agency wants to determine the most applicable predictors of how much its potential clients will spend on overnight vacations so that it can selectively contact those who are likely to spend more. Because no direct data exists, data from a recent Canadian National Travel Survey, filtered to overnight vacations, will be used to evaluate the predictiveness of potential data elements that can be requested from potential clients via a screening form on the travel agency website. The data elements will be chosen based on their predictiveness of total vacation cost as reported in the travel survey, using a regression model. The effect that data elements may have on completion of the screening form and on the potential for the appearance of unfair discrimination will also be considered. Subsequent predictive models built on the selected data elements may be checked against both future travel surveys and the travel agency's own clients.

Task 2 – Outline modeling impacts of data sources and sampling information (7 points)

Quality responses to this question demonstrated knowledge of data collection techniques and how different sampling methods can result in biased data.

Many candidates easily supplied two quality data impacts but struggled providing a third quality data impact. Additional responses which were accepted but not shown here are related to age selection within the household, biasing toward older respondents and differences between response rates with

email and mailed questionnaires. Given the survey nature of the National Travel Survey dataset, care must be taken to ensure it is appropriate to use for predictive modeling:

- If looking at the data on a quarterly basis, Q1 2020 only includes two months of data due to the COVID-19 pandemic, which could cause inconsistency when compared to Q1 in other years.
- Being a voluntary survey, there may be respondent bias in our results, i.e. our data represents only those who chose to respond, a different population than who may contact the travel agency, which may lead to biased results.
- Responses from smaller provinces and lower incomes are given additional weight and are over-represented in the data. This will increase the performance of our models when studying the effects of province of origin and income on travel cost but could cause bias if these factors are ignored.

Task 3 – Explain modeling impacts of high-dimensional and granular data (10 points)

Quality responses to this question demonstrated knowledge of the difference between granularity and dimensionality and the impacts high dimensional data can have on both GLMs and tree-based models.

Many candidates were unable to distinguish between dimensionality and granularity. Nearly all candidates were able to note regional data is more granular than provincial data.

A majority of candidates were able to identify high-dimensional categorical data can cause issues with GLMs. Well-prepared candidates were able to note tree-based methods are also adversely affected by high granularity variables when determining leaf splits.

For a categorical variable, dimensionality refers to the number of different possible values that the variable has. Granularity refers to how precisely a variable is measured. While the dimensionality can always be compared between two variables, the granularity cannot always be compared because they represent different aspects of the same data.

In the destination data, the regional information is more granular than the provincial information.

When modeling with a GLM, high-dimensional categorical variables may result in overfitting when there are few results in some levels. Also, methods for selecting variables in a GLM often include or exclude all levels of a categorical variable at once, not allowing for the possibility that some levels are helpful.

When modeling with a tree-based model, high-dimensional categorical variables continue to pose an overfitting problem. In particular, tree-based methods seek out the optimal subsets of levels of the categorical variable where distinguishing between these subsets produce a better fit, but these groupings are often non-sensical (like combining regions from different provinces) and are difficult to interpret.

Task 4 – Improve the graph (9 points)

Quality responses to this question demonstrated knowledge of effective communication through graphs and an ability to build a graph which effectively communicates travelers by number of companions.

The majority of candidates noted the provided graphs were challenging to interpret in ways similar to those provided below. Well-prepared candidates were able to produce a bar graph to more clearly represent the information. Candidates who provided boxplots or dot plots of single points were not awarded points.

While no credit was given, it was helpful if candidates transferred graphs to the document submitted.

An effective graph of three values should be clearly labeled, including having a title, and should convey the relative size of the three values by displaying them in a form that humans can easily perceive. Using lengths with a common base will convey the values most easily.

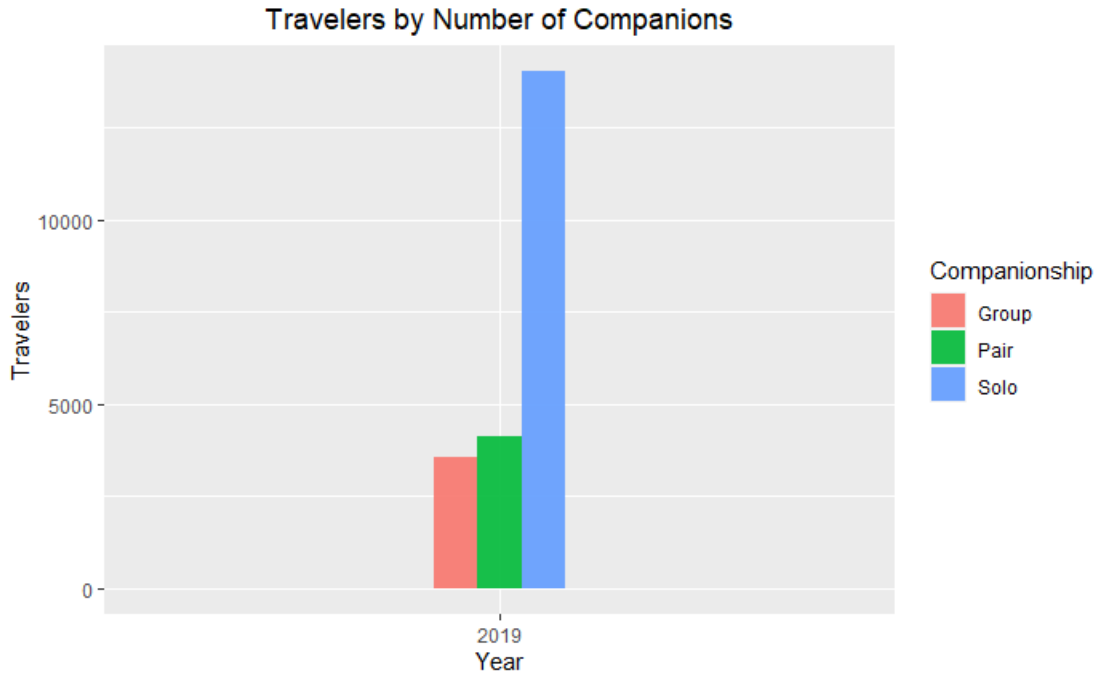
All three graphs lack titles, hindering understanding of exactly what is being viewed.

The first graph's values coincide with lengths but have no common base, which makes it harder to see whether travelling in pairs or a group is more popular in this stacked bar graph. Helpfully, the colors are easy to distinguish, and each is clearly labeled, as are the axes.

The second graph's values coincide with angles, which make perception of relative values even harder than with lengths without a common base. The contrasting colors and labels are helpful, but there is nothing indicating what values the angles are representing.

The third graph is unhelpful except for its clearly labeled axes. Distinguishing hue intensity is difficult and not a good choice when others are available.

The graph below effectively conveys the three values. The title orients the viewer. Clearly labeled axes aid understanding. Using length to express values and a common base make it clear to the viewer that more travelers travel in pairs than in groups.



Task 5 – Filter the data to fit the business problem (7 points)

Quality responses to this question demonstrated an ability to effectively relate data elements to the business problem and identify outliers which are non-intuitive.

For full credit, candidates needed to identify outliers and reasonably justify their removal. Poorly prepared candidates did not take necessary steps to investigate the data to find outliers, and as a result did not remove any outliers. Outliers with respect to long trip length or high trip cost were also accepted.

Candidates generally performed well identifying data elements which could give the appearance of unfair discrimination.

The travel agency offers "overnight vacation planning services," so the 2383 observations where **Reason** = "visit" are removed, leaving only the "vacation" **Reason**. Also, it seems implausible that an overnight vacation would have zero cost, so 18 observations with **Cost** = 0 are also removed, leaving a minimum cost of 10.

Several variables, if requested, would expose the travel agency to the appearance of unfair discrimination. HHI (household income), **Age**, and **Gender** are removed for this reason. Also, the variable Reason is removed because it no longer distinguishes among any of the records and would cause issues when fitting some types of models.

Task 6 – Assess your assistant's hierarchical clustering work (11 points)

Quality responses to this question demonstrated knowledge of how a hierarchal clustering model clusters variables, and how the information is displayed in a dendrogram.

To receive full credit, candidates had to note height of a dendrogram was proportional to similarity and recognize the model was clustering primarily on distance and recommend variable transforms which negated the issue.

Well-prepared candidates identified Distance varies more than Duration, and the best way to handle this disparity is to standardize all variables. Standardizing by just standard deviation and both mean and standard deviation were accepted. Candidates differentiated themselves by how well they communicated the logic behind standardization as it related to the modeling task at hand.

Candidates did not receive points for recommending only one variable be standardized.

Height, as seen when the dendrogram is plotted, indicates how dissimilar two data points or clusters are when they are fused into a single cluster. As there are fewer and fewer clusters, the dissimilarity between clusters increases.

Distance varies far more than **Duration** given their units of measurement. **Distance** ranges from 6 to 4,922 km with standard deviation of 571 km while **Duration** ranges from 1 to 81 nights with standard deviation of 4 nights. The units of km and nights are treated as equals when they likely do not have equal importance. Because two clusters are likely to differ far more by **Distance** than **Duration** when calculating the distance, **Distance** is the dominant factor when determining which clusters to combine.

One approach to balance **Duration** and **Distance** is to standardize each by their mean and standard deviation. The number of standard deviations for each dimension will be far more balanced than the unstandardized data. Another approach is to determine a common unit of value—how many kilometers of **Distance** are equal to an additional night of **Duration**? This equivalence could be assigned by judgment (as 1 km = 1 night is judged as not being appropriate) or determined by a separate model, for example a linear regression on trip cost using **Distance** and **Duration** as predictors.

Assigning a common unit of value is recommended over standardizing by standard deviation. The standardization for most data points can be significantly influenced by a few outlier data points and could vary depending on the dataset. By affixing the relationship of kilometers and nights, distance between points can be kept consistent as the data varies, keeping the clustering result for identifying types of clients more consistent as well.

Task 7 – Explain Differences Between Model Selection Methods (8 points)

Quality responses to this question demonstrated knowledge of different model selection methods.

Most candidates were able to describe each of the three model selection methods well. Exceptional candidates were able to contrast the three methods holistically including the proportion of data used in training and testing, relative degree of overfitting, and how well a model will predict selected using each of the three methods.

Candidates in general were able to recommend one model selection approach with a valid justification using language targeting ageneral audience. Recommending multiple model selection methods for inclusion was also accepted if justified.

The goal of model selection is to compare potential models, having different structures and/or different predictors, and determine which will be most predictive when presented with new data.

With Akaike Information Criterion (AIC), all the available data is used to train (determine the predicted values of) the potential models and the AIC calculates, from this same training data, how well it fit that data, with a preference for models with fewer predictors. Because this method does not directly consider how well the models fit new data, it does not give user direct insight into how well the model generalizes to unseen data. AIC model selection will generally result in a more parsimonious model, and in general more parsimonious models tend to generalize better.

The other two methods, 80%/20% train/test split and 5-fold cross-validation use 80% of the data to train the potential models and 20% of the data to test the models. Because the testing data was not used to train the models, it can be used to directly assess how well each model makes predictions when presented with new data. With the train/test split, this is done only once, and some data is never used to train any model while other data is never used to test any model. With 5-fold cross-validation, the 80%/20% split is done five times on the same data such that all the data is used at some point for both training and testing models, making it less likely that the particular separation into training and testing data influences which model is deemed most predictive.

For choosing which data elements to request from potential customers for the travel agency website, the model selection technique for choosing which variables are predictive is particularly important. Even though it takes longer to carry out, cross-validation is recommended. It does best to reduce overfitting, which will mean selecting fewer variables and having higher confidence that the variables selected will be predictive of travel cost.

Task 8 – Explain Differences Between Weights and Offsets (8 points)

Quality responses to this question demonstrated knowledge of domains and ranges of distributions and link functions as well as knowledge of the difference between weights and offsets.

Many candidates were able to identify a non-negative and right-skewed distribution is needed for modeling the cost in excess of $500. The valid choices include but not limited to Tweedie, Gamma, Inverse Gaussian and Lognormal (or Normal with log link function). Some candidates proposed binomial distribution which is not appropriate as the variable here (i.e. cost in excess of $500) is continuous, not a combination of Bernoulli variables. Exemplary candidates were able to not only identify the proposed distribution needing to be non-negative and right skewed, but the distribution captures a large fraction of observations at exactly zero cost, hence the Tweedie distribution will be the most appropriate distribution to use here.

Most candidates were able to describe what weights and offsets in general are when applied to a GLM, but only few can correctly articulate the specific difference between the two concepts.

Many candidates incorrectly recommend using weights for implementing the assumption that the cost is directly proportional to the number of nights.

Candidates who just referenced R manual in their answers would receive no credits.

The cost in excess of $500 has a large portion of observations at exactly zero cost, while observations not at zero have a right-skewed distribution. The Tweedie distribution, a compound distribution which sums a Poisson-distributed number of gamma distributions, fits this situation quite well. Unlike most non-negative distributions, it includes a non-zero discrete probability at 0, with the density function above 0 being right skewed due to being a positive sum of gamma distributions, which themselves are right-skewed.

When applied to a GLM, weights give unequal importance to the observations for determining model fit based on the impact each observation should have while offsets adjust the prediction for each observation but not its relative importance when determining model fit. In other words, offsets act as a known coefficient for a parameter, rather than a coefficient to be fitted.

For modeling cost assuming it is directly proportional to the number of nights with a GLM with log link function, using an offset is more appropriate. The log number of nights as a term in the linear equation being fitted will have the effect of multiplying the result by the number of nights when predicting the cost.

Task 9 – Bias Variance Tradeoff (4 points)

Quality responses to this question demonstrated how elastic net regression leverages bias variance tradeoff to improve predictive power.

Many candidates were able to broadly describe bias-variance tradeoff, but only a smaller subset could explain the approach can lead to enhanced model performance as the reduction in variance more than offset the added bias. Well-prepared candidates described the elastic net regression shrinks coefficients by adding a penalty term but failed to mention any bias-variancetradeoff and how increasing bias will lead to decreasing variance resulting in improved model performance.

Candidates did not receive credit if they merely mentioned elastic net regression reduced overfitting without providing further explanations.

Elastic net regression increases bias by adding a regularization term to the optimization function, effectively shifting the coefficients closer to zero. Adding bias in this way can be helpful if the reduction in variance more than compensates for the added bias, leading to an overall reduction in model error.

Task 10 – Elastic Net Regression (10 points)

Quality responses to this question demonstrated an ability to interpret inputs and outputs from an elastic net regression model.

Candidates in general can identify from the code that lambda = 0, but some failed to mention lambda determines the size of penalty term. Moreover, candidates in general were able to

determine the best value for alpha, and similar to the first part, some failed to state that the alpha with the lowest deviance would generate the best model fit.

*Candidates in general were able to interpret the predicted impact of **Distance** on the target variable **BigCost,** ranging from a simpler answer that only includes a directional impact to a more comprehensive answer that describes what the specific impact is.*

Points were deduced if candidates did not realize the impact of choosing lambda = 0.

Points were deducted if the alpha corresponding to the lowest deviance was not chosen.

The assistant used lambda = 0. Since lambda is the hyperparameter that controls the size of the regularization penalty, this means there was no penalty and no shrinking of coefficients. Also, standardization was turned off, keeping the data on the same scale as the original. These two items make the models virtually equivalent.

The test metric of deviance is as follows for varying levels of alpha:

alpha	test_deviance
0.00	603.45
0.25	605.71
0.50	605.77
0.75	605.72
1.00	605.84

Lower deviance (being a multiple of negative loglikehood) represents a better fit of the model to test data, so alpha of 0 is the best value of alpha among those tested.

The raw coefficient for **Distance** is 0.001876. As the logit link function associated with the binomial family was used for fitting the GLM, each additional kilometer of distance equates to a 0.001876 in log odds. A more intuitive interpretation is that for each log(2)/0.001876 = 369 km of **Distance**, the odds of **BigCost** (that **Cost** is at least $500) doubles.

Task 11 – Random Forests (8 points)

Quality responses to this question demonstrated knowledge of random forests and the impacts of different parameters and select a model based on AUC.

Candidates in general performed well in this task. The goal for this task is to test the candidates what the different settings in random forest are and how tuning these parameters affect model performance. Candidates should use their own words explaining what these settings mean and how they affect the structure of the random forests as well as their respective difference.

Candidates who just referenced the R help pages in their answers received no credits.

In rf_model_1, each tree in the random forest is fit to 63.2% of the data, sampled without replacement, using floor(sqrt(6)) = 2 randomly chosen predictors. In rf_model_2, each tree is fit to the same number of observations as the data but is sampled with replacement, i.e. bootstrapped, using all 6 available

predictors. In rf_model_3, each tree is fit to the bootstrapped observations for rf_model_2 using 2 randomly chosen predictors as described in rf_model_1. The variations in sampling and selection of predictors will cause the amount of overfitting to vary.

In this classification problem, the models determine their predictions by counting the votes for each individual tree and choosing the category with the most votes. Ties are broken randomly.

The AUC results are as follows:

Model	AUC
rf_model_1	0.6670
rf_model_2	0.6619
rf_model_3	0.6608

The settings for rf_model_1 are recommended because it produced the highest AUC, indicating a better prediction of costs of at least $500 compared to the other two sets of settings. These settings had the smallest data and number of predictors and may have done more to prevent overfitting.

Task 12 – Recommend Data Elements to Collect (13 points)

Quality responses to this question demonstrated knowledge of how to interpret a variable importance plot or table and make a recommendation on data elements to collect based on the variable importance plot or table.

Many candidates were able to display the variable importance data. Additionally, many candidates could somewhat explain what the variable importance results represent. Well-prepared candidates were able to describe how the importance of a specific variable is determined from the random forest. Candidates in general were able to recommend data elements the travel agency should collect either based on better model performance or on some business considerations, but very few are able to provide justification from both aspects.

Full credit was given to candidates regardless of whether a variable importance plot or table was supplied.

Variable	**Relative Importance**
Distance Traveled	174
Number of Nights	76
Province of Origin	59
Number of Others	41
Calendar Quarter	36
Mode of Travel	16

Variable importance is a relative measure that indicates how often a variable was found to be useful in predicting whether travelers spent at least $500 on an overnight vacation. Variables with similar importance are about equally useful while variables with very different importance indicate that the one

with higher importance is much more useful than the one with lower importance. They are determined by observing, in the particular model used, how much making a distinction between groups of people using a particular variable distinguished whether those people would spend at least $500.

Multiple models were run including just the most important variable, the two most important variables, and so on, checking to see when adding more variables failed to improve the prediction of big spenders when presented with new data. For the model validation metric used, area under the curve, a higher value indicates a more predictive model.

Distance Traveled	**Number of Nights**	**Province of Origin**	**Number of Others**	**Calendar Quarter**	**Mode of Travel**	**Area Under the Curve**
Included	Excluded	Excluded	Excluded	Excluded	Excluded	0.6043
Included	Included	Excluded	Excluded	Excluded	Excluded	0.6640
Included	**Included**	**Included**	**Excluded**	**Excluded**	**Excluded**	**0.6897**
Included	Included	Included	Included	Excluded	Excluded	0.6767
Included	Included	Included	Included	Included	Included	0.6670

The travel agency should collect distance traveled, the number of nights, and the province of origin. As each of these were included, the predictive power of big spenders increased. When the next most important variable, number of others, was added, the predictive power was poorer, and including all the variables did not further improve the model as it was finding too many spurious patterns in the data when given the additional flexibility of more predictors.

The distance traveled and province of origin were already planned to be collected. The number of nights should also be freely given by prospective clients, as this is a natural question directly pertaining to travel planning. Other potential variables such as age, gender, and household income were not considered in the modeling, even though they may have improved predictive power, because they could expose the travel agency to reputational risk of unfair discrimination among prospective clients.