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.

Exam PA June 22, 2021 Project Report Template

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 22 attached.

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

Task 1 – Assess the data sources (12 points)

Quality responses to this question included the identification of different data types and an ability to communicate reasoning behind why certain data items may or may not be useful for the modeling problem.

A majority of candidates were able to identify data sources as structured, semi-structured or unstructured.

Most candidates were able to identify reasons why or why not a certain data element would or would not be helpful, with the most popular response highlighting ethical concerns. Many candidates failed to recognize audio recordings had come from the claims process.

Candidates failed to earn points if justification for a data source seemed boilerplate or was nonsensical.

- **Audio recordings**: unstructured, because they are stored as audio files, which doesn't fit in a tabular structure
- **Social media profiles**: semi-structured, because there are various elements that could be stored in tabular format (user name, location, etc.) but some elements like photos cannot be fully represented in tables
- **Demographic information**: structured, as it is stored in a tabular format with data stored as numerical or categorical values
- **Expected trip itinerary**: semi-structured, because while the data could be stored in a tabular format (date, destination, etc.), some elements will involve detailed text that would need to be parsed prior to analytics work

---

To the manager:

In order to enhance our predictive modeling supporting trip cancellation insurance, we have looked at incorporating audio recordings from our call center, social media profiles and images, policyholder

demographic information, and policyholder trip itinerary information into our analysis. The following outlines tradeoffs and considerations for using these data sources:

Audio recordings from call center

Our call center is currently only accessible for existing policyholders, and most calls are regarding our untimely claim payments. For the purpose of predicting spend, the call center audio data is unlikely to add much beyond demographic data and trip itinerary data. In addition, it will be challenging to leverage as it requires a lot of data manipulation to be useable.

Publicly available social media profiles and pictures

There are ethical concerns regarding using social media information, even if it is publicly available. Industry standards classify names and photographs as personally identifiable information so additional care and consideration would need to be given to use this data. Also, from a data manipulation perspective, it would be difficult to access this information in a format useable by a predictive model.

Demographic information on the policyholder

This information is available in existing ABC databases so we can access it with minimal data manipulation. Using it on a de-identified basis also does not pose high privacy risks.

Trip itinerary information

This information is currently stored in raw text format so it would require some data manipulation to parse and, given the nature of free form text inputs, could potentially have some quality issues. However, understanding where members travel today could potentially provide modeling benefits when predicting travel costs.

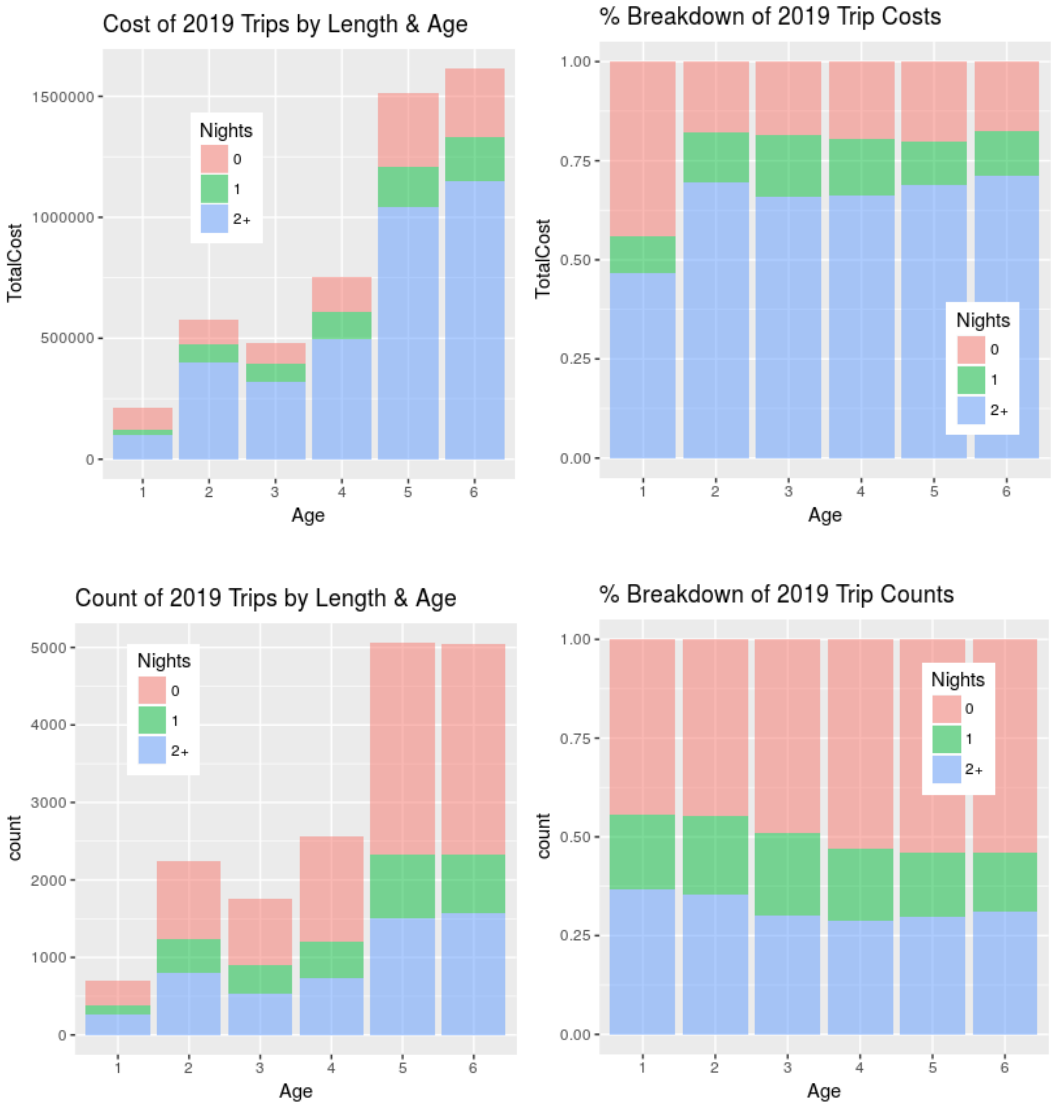
Task 2 – Interpret the graphs (12 points)

Quality responses to this question demonstrated knowledge of how to interpret graphical information.

Many candidates were able to find two conclusions but struggled with a third conclusion. Many candidates noted the discrepancy between breakdown of trip costs and trip counts for the youngest age bracket.

Candidates did not receive full credit if they provided three conclusions, but conclusions were restatements of one another. No candidates provided any rough mental math in the thoroughness presented in conclusion three.

While not necessary for credit, it was helpful if candidates pasted graphs from the project statement into their responses.



- **Conclusion 1:** Although all age groups take about the same proportion of one-day, two-day, and longer trips, adults under age 25 spend substantially more on one-day trips compared to overnight trips than do the other age groups. The bottom right graph shows the steady proportions of trip lengths across all age groups, with about half of trips taken being one-day trips. But in the top right graph, adults in age groups 2-6 (age 25 and up) spend about 20% of their travel costs on these one-day trips while age group 1 (ages 19-24) spend about 40%, as seen in the topmost red portions. These graphs by themselves do not clarify whether their one-day trips are more expensive or their overnight trips are cheaper than those for other age groups, just that they are more similar in price per trip for this youngest age group.
- **Conclusion 2:** The two oldest age groups, age 55 and up, take more trips and spend more money on trips compared to other age groups, assuming that each age group is represented proportionately in the data. The bottom left graph supports the more trips aspect and the top

left graph supports the more spending aspect. In each, the two rightmost bars are at least twice as tall as the other four bars, where it is not expected that, for instance, the 55-64 age group is more than twice as large as the 45-54 age group. It is more believable that they take more trips.

- **Conclusion 3:** Although the two oldest age groups take more trips and spend more money on trips, as noted above, the average cost across all lengths of trips appears not to vary greatly by age group. The comparison of the top left and bottom left graphs support this conclusion, though it is more difficult to see visually since it relies on comparing the ratios of lengths of bars in the top graph to the lengths of the corresponding bars in the bottom graph. Still, since the heights of bars within each graph exhibit similar increases and decreases, moving from one age group to the next, it is fair to conclude that the average cost per trip is similar. Mental division of the top figures to the bottom figures suggests an average of around $300 per trip when ignoring the number of nights, e.g. ~$1.5M on 5000 trips for the two oldest age groups on the right of each graph.

Task 3 – Explain work on the Cost variable (12 points)

Quality responses to this question demonstrated knowledge of how to properly display information in graphical form and how target variable transforms would affect the objective function in either a GLM or a tree-based model.

Most candidates were able to identify issues with the assistant's graphs as well as justify the first and second edits. Candidates failed to earn points if they did not to recognize adding 1 was used to make log transforms possible.

Many candidates identified issues which a right skewed variable would have on a GLM including selecting a right skewed distribution function.

Many candidates failed to receive full credit because they did not recognize a right skewed variable would also impact the objective function of a decision tree, giving less granularity to predictions in the right tail section of the distribution, the part with which the client is most concerned.

The assistant's code gives two graphs lacking titles. The first graph, a scatterplot of **Cost** by **Duration** with a linear trend connecting the two, is helpful for getting a sense of the relationship between these two variables but is hard to verify. The small circles overlap each other without transparency on the relatively few values for **Duration**, obscuring the actual density of points. The linear trend may be heavily influenced by outliers, but at the same time seven outliers were removed in this graph—it would be helpful to see where these outliers are and what the trend line would be with these included, as that will better mimic the modeling outcome with the linear model.

The second graph is a density plot of **Cost** with most values close to zero, but it is hard to discern the shape of the density function due to the inclusion of outliers. The graph does make clear that **Cost** is right-skewed, but one cannot judge from this graph what type of skewed distribution might be appropriate.

After getting a full view of the outliers on the scatterplot, the first edit to the graph removes the two trips costing over $15K, as there is a large gap from trips costing $12K to one costing about $20K. Nine

trips lasting over a month are also removed due to their rarity and separation from other data points, making them more likely to skew certain types of predictive models.

The second edit, adding 1 to all costs, will make little difference to the target but makes possible the use of log scales for visualization, allowing the amount of right skew to be studied more carefully. A log scale is helpful because costs and other economic variables often have proportional or geometric comparisons, and it is easier to visually compare arithmetic (length) comparisons than geometric (proportional) ones.

Cost is right skewed, but only about as much as a lognormal distribution as shown by the final graph. For a decision tree, a skewed target can put heavy emphasis on minimizing the squared errors of relatively few extreme points, causing splits in the tree to find more discrimination in the skewed area than where most of the data is. For a GLM, a distribution that is right skewed should be used to measure the error of its predictions when optimizing the coefficients, or it will strongly fit the right tail of the experience while fitting the majority of observations less well. However, strongly fitting the right tail may be desirable for this high-end product.

Also, **Cost** includes many (formerly) zero-cost trips, as shown by the local mode on the left end of the density plot of log(**Cost** + 1). These are somewhat puzzling and may represent issues in the data but they could be reasonable for comparison purposes. For a decision tree, the concentration of data at the lowest value poses relatively few modeling issues, though many observations at one value can cause overfitting when certain characteristics happen to show up for that one value in the training data. For a GLM, many distributions do not handle a discrete density at the left-most end well, though the Tweedie distribution does and should be considered if the version of **Cost** including zeros is restored. Placing too much emphasis on zero costs in modeling this high-end product is not advisable, however.

Task 4 – Consider regression trees (7 points)

Quality responses to this question demonstrated knowledge of how the objective function of decision trees interacts with right-skewed target variables.

The best responses noted the difference in trees being due to the log transform of the target variable and noticed Tree 2 had more density on the right tail leaves, connecting the discrepancy to the business problem.

Many candidates noted the differences between a log-scaled target variable and a non-transformed target variable but did not go much deeper than surface level conclusions.

Candidates failed to earn points if they did not identify the client is more concerned with the right tail of the distribution and Tree 2 provides much more insight into the right tail than Tree 1.

Tree 1 used the original **Cost** because the values in the top of each leaf look like costs, and those in Tree 2 are smaller and thus use $\log_{10}$**Cost**.

The distribution of the eight leaves in each graph are quite different. In Tree 1, the leaf for the smallest costs covers over half of the data, and five leaves have assigned values over $1,000. By contrast, Tree 2 only has a quarter of the data covered by the leaf for the smallest costs, and only one leaf has an assigned value over $1,000 (whose log is 3). Since the creation of leaves involved minimizing squared differences of actual and assigned values, Tree 1's original **Cost**, having a long right tail, produces more

high valued leaves than Tree 2's $\log_{10}$**Cost**, which has a more bell-shaped distribution. The variables used to make the two trees differ somewhat in the order of splits but the same three variables are used multiple times in each, suggesting that **Distance**, **Duration**, and **Reason** are the most important predictors of **Cost**.

In this business problem, the focus is on a high-end trip cancellation insurance product applicable for pre-paid costs of at least $1000. Tree 2 have seven leaves distinguishing trip costs under this threshold, which is not helpful to ABC. Tree 1, with the untransformed **Cost**, is preferable because it provides multiple branches leading to qualifying trips and distinguishing among these trips, also important to ABC.

Task 5 - Explore Correlations and PCA (8 points)

Quality responses to this question demonstrated knowledge of the correlation matrix, collinearity, and PCA.

The best responses included assessments of at least two of the choices made by the assistant for correlation matrix and PCA, provided an insightful discussion of the correlations observed in the correlation matrix and their implications for modeling, and discussed the PCA results. Correlation observations could have either focused on correlations among the predictor variables and any related collinearity concerns, or focused on correlation with the target variable and implications for a variable's value in the model. Further, they identified that the Principal Components could not be used in predictive modeling, since the target was included in the PCA.

Many candidates failed to receive full credit as they did not include assessments of the choices made for the PCA and correlation matrix.

Many candidates failed to receive full credit as they did not identify that PCs developed using the target could not be used as variables in the predictive model. Additionally, many candidates failed to receive full credit as they did not address the implications of PCA for predictive modeling more generally.

Both the correlation matrix and PCA are performed solely on numeric variables. Interdependency involving non-numeric variables can also cause modeling issues, but these issues cannot be exposed with these two techniques applied to numeric variables unless the categorical variables were transformed into dummy variables, which is worth considering.

Included among the numeric variables is **Age**, which is really an ordered categorical variable. Its inclusion still results in meaningful comparisons with the other variables because it has a natural order and is mostly evenly spaced, though the smaller age group 1 and larger age group 6 (in terms of age range) may bias the two techniques compared to being able to include the underlying age. It may be better to run PCA without **Age** to see what effect it has on the rotations of the true numeric variables.

Cost, the target variable, has moderately strong correlations, just over 0.50, with both **Distance** and **Duration**. This suggests, particularly for linear regression and associated models, that these two variables will be good predictors for the target and should be strongly considered during model selection.

The near 0.50 correlation between **Distance** and **Duration** suggests that collinearity may be an issue when including these as predictors. That these have the same signs in the first three principal components highlights their collinearity in the context of other variables.

In order to fully consider PCA to create input for a GLM or other model, we would need to rerun the PCA without the target variable. Otherwise, we would be modeling assuming knowledge of the target, which we will not have in the future (or else we wouldn't be trying to predict it!). Setting that aside, in the PCA, the cumulative proportion of variance explained rises only modestly quickly and the scree plot does not have that sharp an elbow, suggesting that substituting these components for the underlying variables will not produce much dimension reduction, not enough to offset the considerable difficulties for interpretation its use would create. PCA should be abandoned for this business problem.

Task 6 - Discuss Dimension Reduction Techniques (8 points)

Quality responses to this question demonstrated more detailed knowledge of PCA mechanics for feature generation, selection of principle components for modeling, and how LASSO regularization can shrink coefficients to 0.

Many candidates failed to receive full credit as they did not provide a quality comparison of PCA and LASSO on how they perform dimension reduction. Comparisons of PCA and LASSO unrelated to dimension reduction did not receive credit.

Many candidates discussed interpretability of one or both models, but they failed to receive full credit as they did not compare the interpretability of the models (e.g., they only included separate statements saying each was difficult to interpret).

PCA can be used to create principal components, and these can be used as predictors in place of the original variables. Each principal component is orthogonal (as different in direction as possible) to every other principal component, and typically relatively few components are used in place of many original variables, to the extent that the principal components capture a high proportion of the variance of the original variables. This substitution, specifically the score produced by the linear combination of original variables each component represents, can reduce dimensionality and improve the predictive power of the resulting model.

Where PCA reduces the dimensionality without reference to the target variable, LASSO reduces the dimensionality by coercing some coefficients to zero when optimizing with its penalized optimization function that considers both the fit of the coefficients in predicting the target and the number of dimensions in the model. The reduction in the number of dimensions can be directly chosen with PCA but is only applicable to numerical variables. The reduction in the number of dimensions in LASSO can be only indirectly chosen via the lambda hyperparameter but is also applicable to categorical variables after binarization.

The principal components from PCA can be difficult to describe in detail compared to the original variables, which can be preserved in LASSO.

Task 7 – Consider two transformations of the Age variable (9 points)

Quality responses to this question demonstrated knowledge of modeling impacts of transforming a predictor variable.

Full credit was awarded for identifying binning based on average age would result in a predictor monotonically increasing or decreasing with age. Binning age as factor variable would allow for a non-linear impact of age to fit the training data better but could potentially increase bias.

Candidates failed to earn points if they did not to recognize either of the above two points.

As a factor variable, **Age** would potentially have a different impact on **Cost** for each age band, adding a total of five degrees of freedom to the GLM. The impact from each age band would be independent from the other impacts, allowing any relationship between the age bands and travel costs to be fit. Using the expected average age as a numeric predictor instead would add only one degree of freedom to the GLM. The impact from age would be monotonic across all age bands.

The factor variable version would have higher variance and lower bias compared to the numeric version using expected average age. The additional degrees of freedom allow the factor variable version to fit the training data more precisely and capture a wider range of patterns in the data (low bias), but that also makes it susceptible to overfitting random occurrences in the training data and not generalizing well to unseen data (high variance).

The factor variable version has the advantage of being able to fit non-linear relationships between **Age** and **Cost** with relative ease and will not extrapolate a strong relationship in some age bands as being a uniform relationship among all age bands. The numeric version, in addition to being less prone to overfitting, has the advantage of being easier to interpret and can be converted into a simple rule of thumb.

Task 8 – Consider Two Models and Their Respective Targets (8 points)

Quality responses to this question demonstrated the ability to relate modeling results to the business problem. They discussed that the classification model more directly addressed the potential buyers, but the GLM provided additional information that may inform which customers would buy the most expensive trips.

Candidates failed to receive full credit when they compared advantages and disadvantages of the models themselves (e.g., difficulties in the modeling process) that were unrelated to the differences in model results (e.g., target variable) and their impacts on the business problem.

Both the regression model, predicting travel costs, and the classification model, predicting whether costs are at least $1000, can give insight into factors related to the amount spent on trips.

Although ABC's trip cancellation product is sold to people with $1000 or more of prepaid trip expenses, it is hard to know what portion of total trip expenses reported in the survey data correspond to $1000 of prepaid travel expenses. Most likely, there is not a one-to-one correspondence between the two values. That the data is based on surveys also introduced the problem of faulty recall, and bias in this recall harms a classification model more than it does a regression model. If a classification model is used, choosing a cutoff will introduce additional error into the model, particularly as there are not naturally two distinct groups (e.g., people who take $800 trips and people who take $1200 trips) but rather a continuous distribution of trip costs.

Much of the information in the **Cost** variable is lost when it is reduced to being merely above or below one value as in a classification model, whereas a regression model retains this information, which is potentially valuable for understanding what sort of people spend more on trips. However, the regression model will assume that the marginal effect of each predictor will have a similar effect at lower costs and higher costs. With many of the trips in the data costing much less than $1000, the regression model may do a better job distinguishing the numerous cheaper trips than it does distinguishing the rarer more expensive trips, which is the target ABC is interested in.

Task 9 – Explain the Problem with Unbalanced Classes (5 points)

Quality responses tended to be straightforward and demonstrated knowledge of unbalanced classes and their impacts on specificity, sensitivity, accuracy, and AUC. They related the impact in this instance on specificity to the business problem (marketing's desire to correctly identify those who will take the High Cost trip).

Many candidates failed to receive full credit as they did not address the relation to the business problem.

A few candidates failed to receive full credit as they discussed unbalanced data, but did not explain how unbalanced data leads to a poor result.

Confusion matrix on test data

	Actual 0	Actual 1
Predicted 0	2658	204
Predicted 1	24	110

Accuracy and AUC are obscuring the goal of the predictive model. Accuracy is a weighted average of accurately predicting the low trip costs (specificity) and accurately predicting the high trip costs (sensitivity). On the test data, the specificity is 99% and the sensitivity is 35%. The weights applied to these two measures are the amount of data in each class, and because the classes are unbalanced, most of the weight is on the low trip costs, arriving at 92% accuracy. If everyone were predicted to be low trip costs, the accuracy would still be 90% on this test data, and this model leans too much towards that extreme. It does not reflect that identifying high trip costs is more important to ABC—missing 65% of high trip costs is more harmful to informing marketing about the trip cancellation insurance product than including 1% of the low trip costs.

AUC has the same type of flaw—while it looks separately at sensitivity and specificity for a variety of cutoff points (rather than the assumed 0.5 cutoff for the confusion matrix above), it effectively puts a 90% weight on specificity, of less interest to ABC, and a 10% weight on sensitivity, of more interest to ABC.

Task 10 – Discuss Undersampling and Oversampling (5 points)

Quality responses demonstrated knowledge of oversampling and undersampling and how each results in balanced classes. These responses discussed the increase in sensitivity and the decrease in specificity from either method.

Most candidates were able to explain undersampling and oversampling. Some candidates failed to receive full credit as they did not explain how the techniques would impact the predictions.

While the model solution gives an example of oversampling by sampling with replacement, including multiple copies of the minority class was also accepted.

Undersampling keeps all instances of the minority class (like high trip costs in this case) and samples from the majority class (low trip costs). Oversampling keeps all instances of the majority class and samples with replacement instances of the minority class so that there are more minority class instances than there were previously. In both techniques, the imbalance between the majority and minority classes is made less severe, allowing many modeling techniques (including GLM's) to pick up the signal of the minority class more reliably. Because the balance of observations is shifted to increase the prevalence of the minority class, the predicted probabilities increase for the minority class and decrease for the majority class. Given the same cutoff for converting a predicted probability to a predicted class before and after applying these techniques, in the training data the minority class is predicted more often after applying undersampling or oversampling.

Task 11 – Implement Oversampling and Explain the Confusion Matrix (14 points)

Quality responses accurately identified that oversampling before setting aside the test set would result in "cheating," as there would be duplicate records in the train and test data sets. They compared the original and oversampled confusion matrices for the assistant, noting the changes in specificity, sensitivity, accuracy, and balanced accuracy. They separately wrote, in non-technical language, an explanation for their manager of the oversampled confusion matrix, discussing the true positive, true negative, false positive, and false negative rates and related the true positive rate to the business problem. They requested, in non-technical terms, information from marketing on the profit expected from a true positive and the cost of a false positive, so that they could better determine a cutoff value.

Most candidates accurately described why oversampling must be applied after splitting into train and test data. Some candidates failed to receive full credit as they incorrectly thought that the need to apply oversampling after the train/test split was limited to the use of a stratified test set.

Candidates generally did well describing the original and oversampled model for the assistant, although most failed to receive full credit as they did not address all of specificity, sensitivity, accuracy, and balanced accuracy.

Many candidates failed to receive full credit as they did not provide insight on how oversampling is helpful for the business problem.

Some candidates failed to receive full credit as they tried to interpret the model coefficients instead of the confusion matrix.

Again, many candidates failed to receive full credit as they did not discuss the full confusion matrix for their manager.

Many candidates failed to receive full credit as they missed that the information from marketing was to help inform the cutoff value and had very general requests, such as for more data.

Candidates that did poorly explaining the confusion matrix for the manager almost always used the same technical language that they used for the comparison for the assistant.

Some candidate answers appeared rushed with this last response, and would have benefitted from additional pacing.

It was not clear for some candidates when they were trying to provide the information for the assistant and when they were trying to provide the information for their manager. In contrast, some candidates included intros like "To the manager:", "Dear Assistant" or "Dear Manager" to make this very explicit.

Oversampling should be performed after splitting train and test data and only to the train data, because otherwise the resampled minority class observations that get duplicated could end up appearing in both the train and test data, breaking the definition that test data is unseen data and effectively allowing the model to memorize the correct answers for the duplicated records.

Metrics Comparison on Test Data

	Original Model	Oversampled Model
Sensitivity	35%	76%
Specificity	99%	88%
Balanced Accuracy	67%	82%

After oversampling, the sensitivity, measuring how often actual high trip costs are predicted as such, has increased substantially because the probabilities of the minority class were increased through oversampling and the cutoff of 0.5 was not changed. At the same time, the specificity, measuring how often actual low trip costs are predicted as such, has decreased for the same reason. The balanced accuracy metric, which applies equal weight to specificity and sensitivity, increases with oversampling, where accuracy (not shown) decreases (below that of predicting all trips being low cost) because it puts weights proportional to the unbalanced majority and minority classes on specificity and sensitivity respectively.---

To the manager:

Below you will find a confusion matrix, a discussion on how to interpret the model fit by way of confusion matrix and a recommendation on which additional variables the marketing department should collect.

Confusion Matrix on Test Data from Oversampled Model

	Actual < $1000	Actual $1000+	Total Predicted	% of Predicted Correct
Predicted < $1000	**2370**	**75**	**2445**	**97%**
Predicted $1000+	**312**	**239**	**551**	**43%**
Total Actual	**2682**	**314**	**2996**	
% of Actual Correct	**88%**	**76%**		

The above table, called a confusion matrix, displays the results of applying our model to predict trip costs of at least $1000 to data it has never seen before, the test data. The green numbers in the top left

portion of the table are how many predictions were correct, and the red numbers are how many predictions were not correct. These are summed for both actual and predicted results, and the percentage of correct results for each is calculated.

Of the 314 actual high-cost trips, assumed to qualify for our trip cancellation insurance product, 239 of them are correctly predicted as high-cost trips. This 76% success rate for the positive event we are trying to predict is called model sensitivity. If we used this model for targeting our marketing, we would reach about three-quarters of our potential buyers and miss about one-quarter of them.

Of the 2682 actual low-cost trips, assumed not to qualify for our product, 2370 are correctly predicted as low-cost trips. This 88% success rate for the negative event is called model specificity. By keeping this figure high, we reduce how often those who are not our potential buyers receive our marketing.

There are 312 false positives, where marketing would target individuals not expected to be able to use our product, and 75 false negatives, where we would not target individual who could use our product. If we knew more about the relative costs of the unnecessary marketing and missed opportunities, we could refine the predictions of the model to further optimize the cost-benefit tradeoff of our market targeting model.