BrainStation
1-800-903-5159

Data Science Diploma Program

Student Transcript

A minimum passing grade of 70% is required for completion of the Data Science Diploma Program.

Units	Weight	Grade
Unit 1: Data Fundamentals	**15%**	92.5%
Unit 2: Analysis for Data Science	**20%**	90.2%
Unit 3: Machine Learning Techniques	**20%**	92.5%
Unit 4: Big Data Fundamentals	**15%**	90.5%
Unit 5: Professional Development	**30%**	87.9%
Total	100%	

Cumulative Grade: 90.3%

Student ID

Student ID:	571333
Jason	Taylor

Data Science Diploma Program	
Start Date:	January 23, 2023
End Date:	April 14, 2023

Program Completion	
Status:	Diploma Awarded
Transcript Issued:	19-Apr-2023
Withdrawal Date (if Applicable)	

Attendance Grade

Below is a weekly breakdown of your attendance record. A minimum attendance grade of 90% is required for completion of the Data Science Diploma program. Please notify your TA of your absence with a reason ahead of time.

Week	Mon	Tue	Wed	Thur	Fri	Grade
1		7.0	6.5	7.0	7.0	98.2%
2	7.0	7.0	7.0	7.0	7.0	100.0%
3	7.0	7.0	7.0	7.0	7.0	100.0%
4	7.0	7.0	7.0	7.0	7.0	100.0%
5		7.0	7.0	7.0	7.0	100.0%
6	7.0	7.0	7.0	7.0	7.0	100.0%
7	7.0	7.0	7.0	7.0	7.0	100.0%
8	7.0	7.0	7.0	7.0	7.0	100.0%
9	7.0	7.0	7.0	7.0	4.0	91.4%
10	7.0	7.0	6.0	7.0	7.0	97.1%
11	7.0	7.0	7.0	7.0		100.0%
12		0.0	7.0	7.0	7.0	75.0%

Overall Attendance Grade: 96.8%

Grading System

The BrainStation grading system employs a numerical marking system. Below is a description of grade meanings.

Grade Meanings	Numerical Scale of Marks
Excellent	90-100%
Very Good	80%-90%
Good	70%-80%
Developing	60%-70%
Limited	0-60%

BrainStation
1-800-903-5159

Unit 1: Data Fundamentals 15%

Student ID

Student ID:	571333
Jason	Taylor

Unit Grade	**92.5%**

Deliverable 1 (65% of unit grade)					**Final Grade**	**93.1%**
					Late Submission Penalty	
		Exemplary	Very Good	Satisfactory	Developing	Limited
	Comprehension		X			
	Execution		X			
	Communication		X			

Deliverable 2 (35% of unit grade)					**Final Grade**	**91.5%**
					Late Submission Penalty	
		Exemplary	Very Good	Satisfactory	Developing	Limited
	Comprehension		X			
	Execution		X			
	Communication		X			

Additional Comments

Deliverable 1

Excellent first deliverable Jason, great job! Overall, the work reveals a very good grasp of SQL commands and queries. Great comments in the script to explain your thought process, while adding more information and insights from the query results would be recommended. There is very good communication in the business report including the findings summary and external research to back your reasoning. Also, good work on leveraging visualizations to help readers grasp the core information. However, structuring the reports into smaller segments with subtitles and highlights would definitely help communicate more efficiently. Keep up the good work! Detailed feedback is written below:

Comprehension

You showed a strong understanding of SQL fundamentals, good job!

Q2.1 Good work on trying different methods and comparing the results.

Q6.5 Good job of stating your hypothesis and then conduct the testing.

Execution

Q5.1 Great job on trying to combine both starts and ends, an easier way is to JOIN two tables on the common column, which is 'time_of_day'.

Q6.3 Query runs correctly, but can be further simplified and runtime can be reduced if GROUP BY before the matching, adding station name would be recommended.

Communication

Q1.5 Great line chart to compare the average trips between two years as well as adding the temperature information.

Q2.2 Good visualization to compare the membership percentage, while adding the total number of trips would be better to show the distribution between member and non-member usage.

Q5.2 Good job on the bar chart comparing trips between starting and ending throughout the day.

Q6.5 Using a map to support your hypothesis and reasoning is very good, however, adding titles and more detailed information for your graphs would be even better.

Deliverable 2

Great deliverable, Jason! The tableau workbook showed a solid understanding of tableau as a visualization and dashboarding tool. The dashboard was interactive and well done. The business report has good insights and summary, however there's still room to improve. The following detailed feedback would help you to improve further:

Comprehension

The workbook had good choices of visualizations for each question, with appropriate filters, calculated fields and good caption for majority of the charts.

Q3.2 Since it's assumed that every non-member trip is a single trip, so filtering ONLY Non-Member trip revenue is needed.
Q3.3 Very good use of bar chart and heatmap, however, only non-member(single) trip with 30 minutes or less needs to be included.

Execution

The charts and dashboard are very well done. Calculated fields are built efficiently in most of the questions.

Q1.4 Good use of the top N filter + parameter! It is also nice to have labeled values.

Q2.1 Great scatter plot with trend line! Annotating the outlier stations would be even better.
Q2.2 The histogram can be built simpler by dragging the 'Duration Min' directly to the sheet and chose 'histogram' under 'Show Me', then you could further adjust the 'bin' size.
Q2.3 Great work on distinguishing station duration by color, adding size mark and some annotations would be even better, especially for highlighting the top ones.

Q3.1 The calculated field for 'Trip Length' could be simplified. The first part of the condition (before AND) is not necessary as it's already been filtered by the previous condition.
Q3.3 Choosing 'Entire View' would help present the whole graph on one page.

Q4 The dashboard is very well done in presenting core insights to operations team, and great use of filters and parameters, as well as the highlight of key indicators!

Communication

Good use of caption in Tableau! The business report could be further improved by structuring into sections with clearer and descriptive subtitles, and visualizations could be stronger with necessary titles, annotations legends, and axis labels.

Q1.5 Addressing the questions in the business report too is recommended.

Q2.1 Supporting your insights with your Tableau visualization would be recommended.
Q2.2 Reducing the bin size graph will make the distribution better for comparison.

BrainStation
1-800-903-5159

Unit 2: Analysis for Data Science 20%

Student ID

Student ID:	571333
Jason	Taylor

Unit Grade	**90.2%**

Deliverable 1 (40% of unit grade)

Final Grade	**94.0%**
Late Submission Penalty	

	Exemplary	Very Good	Satisfactory	Developing	Limited
Comprehension		X			
Execution		X			
Communication		X			

Deliverable 2 (60% of unit grade)

Final Grade	**87.6%**
Late Submission Penalty	

	Exemplary	Very Good	Satisfactory	Developing	Limited
Comprehension		X			
Execution		X			
Communication		X			

Additional Comments

Deliverable 1

Excellent deliverable, Jason! The deliverable shows a good command of statistics, pandas and the plotting libraries. The notebook is well structured with good table of contents and dataframe version tracking. Codes are concise, and there were comments in code cells where appropriate, as well as good markdown comments on the results and insights, some of the visualizations could be improved though.

Comprehension

The notebook shows a good understanding of statistics, data processing using pandas and visualizations.

Q1.4 Good catch on address and block, year week and date have the same information based on looking at the data. There are other pair of columns that contain same information such as mosquito ID and species.
Q1.5 Great step by step workflow for analyzing null values, and sanity check after imputations.

Q3.2 Plotting the comparison between the AVG number of mosquitoes caught per trap, a bar chart would be more appropriate, as it's not comparing the distribution or percentage of mosquito captured across four traps.
Q3.3 Good job on plotting the AVG number of mosquitoes per trap, however since the data is capped at 50, so the mosquito numbers at a trap level will not be correct. Understanding the sum of mosquitos across different years would give you a better picture.

Execution

The workflow of the deliverable shows good application of pandas and plotting tools.

Q1.4 Since we cannot check every row, writing the code to do a row by row check is required. Good job on doing a sanity check post deletion.

Q2.1 Great job on checking mosquito number variation with time at different granularities (year, month, day)!

Q3.2 As an alternative consider a box plot that shows median, outliers and quartile range for each trap type. Since GRAVID is the most used trap, it is not surprising that they caught the most mosquitoes. The average mosquitos caught per trap might give us additional information on a trap types effectiveness.

Communication

Overall, communication is comprehensive with good comments in code cells where appropriate as well as good markdown of results and insights. Very good job on the table of contents! Try adding back to top in the notebook as well. Good job on listing different version of the data frame and keep the tracking throughout the notebook!

Q1.2 For one categorical variable, typically, conservative monochromatic or dichromatic color schemes are considered best practice.

Q2.1 The rainbow color is nice but less professional for data from the same group, one color diverging would be better.

Q3.3 Good job on pointing out the relatively low number in 2009, it is important to caution readers about the relatively smaller sample sizes.

Deliverable 2

Very good deliverable, Jason! The deliverable shows a strong grasp of linear and logistic regression, preprocessing, and hypothesis testing. Detailed feedback follows:

Comprehension

The deliverable shows a solid understanding of statistics, hypothesis testing and regression models.

Q2.1 Adding the reason of adopting Chi-Square testing, and checking whether the assumptions of test are met is recommended.

Q3.1 The dependent and independent variables in a regression model do not need to be normally distributed by themselves--only the prediction errors need to be normally distributed. Checking the assumptions for residuals after modeling is required to evaluate the model performance.

Execution

Q1.1 Very good EDA on looking at the dataset at the beginning. And good sanity check after creating dummies.
Q1.2 Good approach on calculating the average and plotting on bar chart, codes could be simpler though, using 'map' to match the value for example.

Q3.1 After modeling, checking the assumptions that whether residuals are normally distributed as well as the homoscedasticity is needed. Correlation heatmap does not need to be plotted again, as the remaining correlations won't be affected if you add or drop other columns.
Q3.2 Good approach on plotting the accuracy curve and great interpretation by identifying the imbalanced dataset. Calculating and interpreting odds ratio to interpret the model is recommended.

Communication

Communication is comprehensive with great insights on model interpretation.

Q2.2 Very good chart of displaying both coefficients and p-values as legend, as well as good interpretation of the chart and comments in markdown.

Q3.1 The hypothesis testing in regression models is to check each variable's coefficient separately to identify which variables are statistically significant, rather than considering them as a whole in terms of relationship with target variable.

BrainStation
1-800-903-5159

Unit 3: Machine Learning Techniques 20%

Student ID

Student ID:	571333
Jason	Taylor

Unit Grade	**92.5%**

Deliverable 1 (40% of unit grade)					**Final Grade**	**90.6%**
					Late Submission Penalty	
		Exemplary	Very Good	Satisfactory	Developing	Limited
	Comprehension		X			
	Execution		X			
	Communication		X			

Deliverable 2 (60% of unit grade)					**Final Grade**	**93.8%**
					Late Submission Penalty	
		Exemplary	Very Good	Satisfactory	Developing	Limited
	Comprehension		X			
	Execution		X			
	Communication		X			

Additional Comments

Deliverable 1

Great job, Jason! Good efforts and work on diving into the non-numeric columns and extracting core information from these columns. The deliverable shows a very good grasp of preprocessing and text analysis, with good visualizations. Detailed feedback is below:

Comprehension

Overall, the deliverable showed great comprehension on EDA and NLP.

Q2.3 Very good steps on checking the non-numeric columns and clear transformation plans.

Q3.2 Good job on converting the address, hotel name, reviewer nationality columns. There is however room to improve on the column 'Tags' by extracting core information, such as trip type, room type etc.

Execution

Notebook has good workflow in terms of preprocessing, while codes could be further simplified.

Q1.1 It's always better to try to impute missing values and keep original information rather than dropping the NaN values. Also, performing a more detailed check on the duplicates is recommended.
Q1.2: using round() by adding a small amount to each number is a better and more efficient approach.
Q1.3: A bar graph is preferable to a histogram for discrete numeric columns.

Q3.2 Good efforts on applying for loops to extract elements and information from categorical features, however, try using more simplified codes such as lambda function to extract information more efficiently. It's not recommended to hard code dictionary, try leveraging existing libraries.
Q3.4 Adding the 'p_' and 'n_' to tokens can reduce any term confusions as the words may be the same but the intent is not, it's better make sure not to treat them as duplicated columns.

Communication

Overall, communication is clear and well structured, with good code comments and markdowns.

Q2.1: The question asks explicitly for which columns are numeric. It requires to infer this based on the data types, the data dictionary, and a sample of their values.

Q3.1 Adding more detailed comment on your distribution plot in terms of target class would be better.
Q3.5: An example of what a min_df of x might do, and calling out that argument can be a percentage, and explaining what document frequency means would make for a more complete answer.

Deliverable 2

Great job, Jason! The deliverable shows an excellent grasp of classical machine learning techniques and hyperparameter optimization. The markdown answers on machine learning theory were easy to follow, It's important to support machine learning methodology with the EDA methods learned in the previous unit, for instance, using visualizations such as box plots to explore new features.The analysis of the new model developed after adding a new feature could be more comprehensive. Overall, well done! Detailed feedback follows:

Comprehension

Overall, very good understanding on the classification models as well as their hyperparameter tuning.

Q2.2 It's also recommended to further look at the prediction time for KNN to better highlight relationship, as scoring is what's computationally expensive.

Q3.2 For DT models, scaling is not necessary which increased the computation time.

Q6.2 The analysis on the model developed after adding the new feature could be more comprehensive, such as looking further at the coefficient (odds ratio) on the newly added feature, which may help you understand better the impact of this feature given the overall score stays the same.

Execution

Good workflow overall, it's better to increase the max_iter in this case to help model converge.

Q1.1 Increasing the max_iter could help model converge.

Q 6.1: Doing some EDA with perhaps a box plot to show separation of good and bad ratings based on the new feature would help back up the hypothesis.

Communication

Good markdown comments with high quality visualizations, however, adding more insights could be even better.

Q2.5 Commenting more on the impact or result of data leakage on model performance, such as increased risk of overfitting and model generalization reduced would be better.

Q3.1 Looking at both train and test score will give you more information on model results.

Q5.1 Adding more detailed comments and insights on the shape/trend of the accuracy curve given different C values would be better.
Q5.2 Good points on the TP and FP numbers as well as mentioning precision and recall, explaining further the difference scores in two classes given imbalanced dataset as well as the default threshold setting would be even better.

BrainStation
1-800-903-5159

Unit 4: Big Data Fundamentals 15%

Student ID

Student ID:	571333
Jason	Taylor

Unit Grade	**90.5%**

					Final Grade	**90.5%**
					Late Submission Penalty	
		Exemplary	Very Good	Satisfactory	Developing	Limited
	Comprehension		X			
	Execution		X			
	Communication		X			

Additional Comments

Deliverable 1

Great job, Jason! The deliverable shows a strong understanding of using Hadoop and Spark for big data wrangling. The assignment questions were executed well with smooth workflows. However, there is room for improvement in the communication aspect, both in the report and notebook, where you could provide more descriptive comments and findings. Feedback is below:

Comprehension

Overall, very good understanding of Big Data Fundamentals, and cloud-based distributed computing environment using Hadoop, Spark.

Q5 The result of the getmerge is on the head node, so should use the ls command without hadoop.

Execution

Q4 Since you have created jupyter notebook under cluster, opening notebook directly would be a faster approach.
Q4.2 After filtering the data, it would be beneficial to conduct a thorough sanity check and describe the new dataset in more detail.
Q4.3 It's better to always add the extension to the file to avoid future misleading, and to perform a check on the HDFS. Good sanity check by reading the filtered data again in the notebook.

Communication

Overall communication could be further improved to add more descriptive comments.

Q4 It's recommended to always have markdowns of summary and findings in the notebook for readability.

Q7 Good job on the plots by converting the y axis to unit of a thousand to increase readability.

Q8.2: Explaining the role of master node, data node and name node would make the answer more comprehensive.

BrainStation
1-800-903-5159

Unit 5: Professional Development 30%

Student ID

Student ID:	571333
Jason	Taylor

Agile Project Planning (10% of unit grade)

			Meets Requirements	Partially Meets Requirements	Does Not Meet Requirements	Days Late
Deliverables	Areas of Interest	25%	X			
	Agile Plan and Data Wireframe	25%	X			
	Progress Presentation	25%	X			
	Final Presentation	25%	X			

Grade: 100.0%

Capstone Report (90% of unit grade) **Performance Rating**

			Exemplary	Very Good	Satisfactory	Developing	Limited	Incomplete	Days Late
Categories	Business Problem Definition	5%	X						
	Data Acquisition, Quality and Completeness	5%	X						
	Preprocessing, Exploration and Analysis	15%	X						
	Visualizations	15%		X					
	Modeling and Model Evaluation	30%			X				
	Communication	30%		X					

Grade: 86.5%

Unit Grade

Unit Grade: 87.9%

Additional Comments

Deliverable 1 - Areas of Interest

Well done thinking through the start to your capstone project! Predicting judicial outcomes based on data is a strong fit for a machine learning project. Machine learning models that can be a good fit for this project include logistic regression, a second supervised model such as random forest, k-nearest neighbor, or support vector machine models, and artificial neural networks (ANNs). To help interpret the patterns and trends of the model outputs, it may also be helpful to run an unsupervised learning K-means clustering model using plotly to graph the decisions as a scatterplot and better understand what the cases have in common. If time permits, you can also incorporate tools we will cover in Unit 3 such as writing pipelines and advanced for loops to simplify writing clean, concise code.

The next steps are to load the dataset files and begin data exploration/ eda. It will also be important to write a clear project sprint plan to help with time management. Keep up the great work!

Deliverable 2 - Sprint Plan/Data Wireframe

Good job drafting your data wireframe and sprint plan. The workback schedule includes a realistic timeline and recognizes that there will be an iterative approach to data cleaning and EDA that will be updated based on model evaluation. For a multi-class prediction project, random forest models, xgboost, and/or artificial neural networks are all good choices to improve the prediction accuracy.

Remember to also include visuals throughout the stages of the project as this is 15% of the marking rubric. There is also the option to include python scripting in addition to Jupyter Notebooks to practice building hidden test cases for your testing set if time allows in Week 11/12. The team is looking forward to reviewing the next steps!

Deliverable 3 - Progress Presentation

Excellent job on the presentation! The structure of the presentation was clear and well-organized, and the motivation was effectively communicated. The target variable engineering was a great example of simplifying a complex problem for the audience, and you showed good awareness of the implications of class imbalance. The slides were straightforward and easy to understand, which is important in effectively communicating complex ideas. It was great to see the challenges highlighted, as this is a crucial part of our work as Data Scientists. The responses to the questions asked were also well thought out and informative. Good luck with the remainder of the Capstone work!

Deliverable 4 - Final Presentation

Great job on the presentation, Jason! The diagrams were creative and captured the attention of the audience. It was great to see the diverse set of models experimented with, and the key features extracted from the logistic regression model were very informative. The next steps outlined were actionable and tangible for anyone who would like to take over or collaborate on the project. The slide and explanation of the target variable classes/thresholds were enjoyable and easy to understand. The polarity trend in results over time was also interesting to see. Lastly, good job highlighting the bias in the data and analysis. Overall, excellent work!

Deliverable 5 - Capstone Submission

Excellent job on the capstone, Jason! You demonstrated great fluency in data wrangling categorical and numeric data with python, and a sound foundation in machine learning techniques. There are opportunities for improvement in the communication, but overall, excellent work. Detailed feedback and suggestions for improvements are below:

Business Problem Definition

- Good explanation of problem statement, commenting more on the potential application and impact of the project would be even better and backing up claims with some statistics will help make the argument even stronger.

Data Acquisition, Quality and Completeness

- Good job on finding a rich dataset for your problem.

Preprocessing, Exploration and Analysis

- The preprocessing was thorough with a lot of sophisticated code to clean and qualify the data. Well done preprocessing the categorical columns (e.g mapping and grouping) and applying lambda functions to fill in missing values, however, there were a few instances of hardcoding that could have been improved upon.

Visualizations

- You did a great job with the visualizations for the cleaning and EDA part. However, there are a few areas where improvements could be made. For instance, in some cases, choosing a more appropriate format to present the data, such as using a stacked bar chart for comparing categorical data, could make the visualizations more effective. Additionally, adding more suitable axis labels to certain visualizations would help readers better understand the graphs. Overall, the visualizations were good, but these small adjustments could further enhance their impact.

Modeling and Model Evaluation

- Great job fitting so many models! Your work shows a sound foundation in machine learning techniques and perseverance in tuning the models for optimal generalizability.
- Excellent job on interpreting the feature importance for logistic regression by comparing odds ratio and p-value. It would also be beneficial to explore the feature importance from tree-based models. This would provide additional insight into the relative importance of different features and further inform model selection.
- When conducting a grid search on cross-validation, it's essential to use pipelines for feature engineering, such as OHE and scaling, to avoid data leakage. Additionally, adding regularization and adopting PCA are effective ways to remove multicollinearity. Incorporating these techniques will improve the robustness and accuracy of the models.

Communication

- The report is well-structured and easy to follow. It would be beneficial to include the business impact of the models in the conclusion section since they effectively verify your hypothesis and provide meaningful results. This would add more value to the report and help readers understand the practical implications of the findings.
- There was close to little code comments and markdown in the some of the notebooks, such as EDA. Given the sophisticated thought processes involved in the cleaning and analysis process, it would be helpful to provide more detailed explanations and interpretations of the methodology and analysis plan. This would make it easier for readers to follow the analysis and understand the decisions made throughout the process.
- In the user defined functions, and in longer blocks of code, consider adding some comments to explain what is happening in the code.