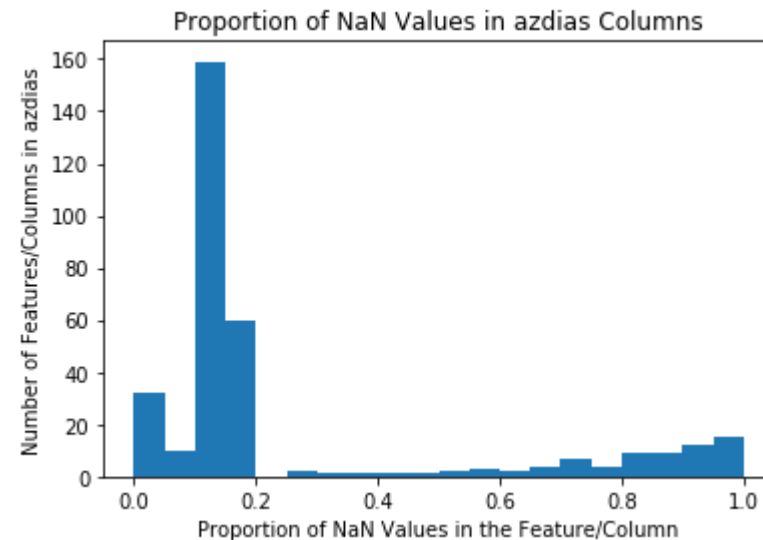

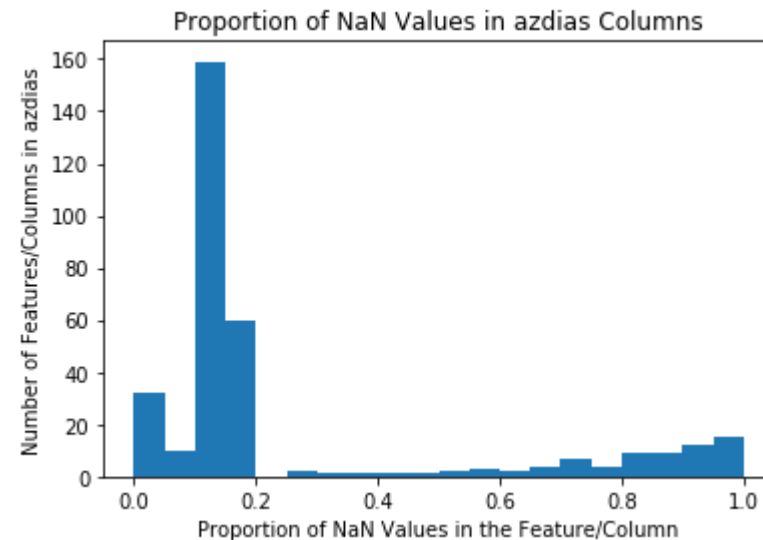
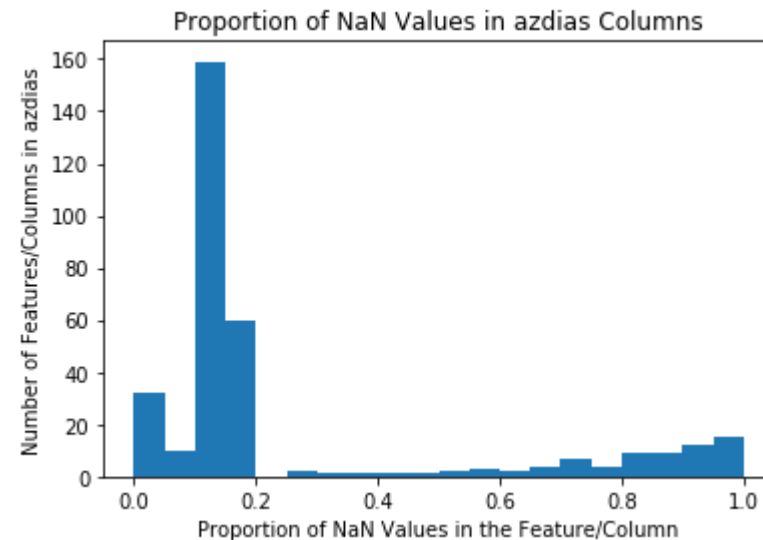


Machine Learning Capstone Project

REVIEW	CODE REVIEW	HISTORY																								
<h2>Meets Specifications</h2> <p>Dear student</p> <p>Great job on your project report! Allow me to be the first to congratulate you on completing the Machine Learning Nano Degree. I feel that you've met or exceeded the specifications for this course and your deeper dive into supervised learning has been quite successful. Again, congratulations on passing and I wish you all the best of luck with your future programming endeavors.</p> <p>Cheers!</p> <h3>Definition</h3> <table><tr><td>✔</td><td><p>Student provides a high-level overview of the project in layman's terms. Background information such as the problem domain, the project origin, and related data sets or input data is given.</p><p>Second, and most importantly, our end-goal is to become sufficiently confident in predicting, based on given demographics data, who (from the general population) would be highly probable to become a new customer of the mail-order company: these individuals would then be targeted by the mail-order company through their marketing campaign. This, in particular, would optimize their marketing action and, ultimately, their customer acquisition process, through informed and sensible customer campaign.</p><p>Nice overview of the problem domain! I love the focus on the real-world impact of the application.</p><p>Suggested:</p><ul style="list-style-type: none">It's a good idea to cite some of the studies where the machine learning techniques that you're using were pioneered. This shows that you really know the field and it gives credit to the inventors.</td></tr><tr><td>✔</td><td><p>The problem which needs to be solved is clearly defined. A strategy for solving the problem, including discussion of the expected solution, has been made.</p><p>The problem can be quantified in the following terms: number of current/established customer and general population clusters (customer segmentation unsupervised problem) and probability of being a new customer to the company (supervised problem).</p><p>You've done a great job restating the problem clearly!</p><p>Suggested:</p><ul style="list-style-type: none">This is a good point to begin to justify why your solution is a good 'fit' for the problem. If you were submitting this to a journal for peer review, you'd want to keep the readers focused on what you want them to think about. If they get distracted, they can ask for random things in subsequent revisions (which can significantly drag out the process and lead to arguments).</td></tr><tr><td>✔</td><td><p>Metrics used to measure the performance of a model or result are clearly defined. Metrics are justified based on the characteristics of the problem.</p><p>Our final supervised model (used to make predictions) is submitted to Kaggle, in the "Udacity+Arvato: Identify Customer Segments" competition, where the evaluation metric used is the AUC for the ROC curve, i.e. the Area Under the Curve for the Receiver Operating Characteristic (ROC) curve. The significant class imbalance in the MAILOUT_TEST dataset motivates us to use this metric, instead of, for example, a basic accuracy score (which would have heavily been impacted by the imbalance, thus not depicting an accurate picture of obtained results and accuracy).</p><p>Excellent justification here! AUC is probably the most optimal metric for this problem. Precision and recall can also be useful for imbalanced datasets where you want to emphasize a model's performance in a specific area. MCC is another commonly used metric for imbalanced multi-class problems.</p></td></tr></table> <h3>Analysis</h3> <table><tr><td>✔</td><td><p>If a dataset is present, features and calculated statistics relevant to the problem have been reported and discussed, along with a sampling of the data. In lieu of a dataset, a thorough description of the input space or input data has been made. Abnormalities or characteristics of the data or input that need to be addressed have been identified.</p><p>Nice job documenting the scope and nature of the dataset for the reader! A sampling of the raw data is provided for the reader and you've noted several unusual properties of the dataset.</p></td></tr><tr><td>✔</td><td><p>Algorithms and techniques used in the project are thoroughly discussed and properly justified based on the characteristics of the problem.</p><p>I think that you've given the reader a really solid overview of the methods and techniques in your project! The emphasis on the unsupervised learning algorithms is excellent.</p><p>**Warning: a spelling error has been made in the Project Notebook file, where SVC has been incorrectly named 'Support Vector Machine' (SVM).</p><p>You probably don't need to explicitly note this for the reader. I'd assume that anyone who was directly reviewing your code would entirely understand what you meant.</p></td></tr><tr><td>✔</td><td><p>A visualization has been provided that summarizes or extracts a relevant characteristic or feature about the dataset or input data with thorough discussion. Visual cues are clearly defined.</p><div><p>Looks good! Your figures are all fully labeled and described in the report text.</p><p>Suggested:</p><ul style="list-style-type: none">It's a good idea to title or name your figures. This will make the report look more polished and it makes it easier to refer to your data in the text (e.g. "see Figure 1...").</div></td></tr><tr><td>✔</td><td><p>Student clearly defines a benchmark result or threshold for comparing performances of solutions obtained.</p><p>As mentioned in the Project Proposal, we chose a basic Logistic Regression Model as a benchmark model, against which we will compare the final trained model in Part 2.</p><p>Sounds great! This is a super common default implementation that shouldn't be too hard to beat.</p></td></tr></table> <h3>Methodology</h3> <table><tr><td>✔</td><td><p>The process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented. Complications that occurred during the coding process are discussed.</p><p>You've done a really nice job documenting your implementation for the reader. I don't think that a skilled programmer would have any difficulty reproducing your results using only the project report (since you've fully documented the model hyperparameters for each model). One other thing that I'd suggest adding here would be to note if there were any complications or difficulties that you encountered during the coding process.</p><p>EDIT: I see you mentioned some issues you encountered with the SVC being slow to train/tune later in the report. Excellent job!</p></td></tr><tr><td>✔</td><td><p>The process of improving upon the algorithms and techniques used is clearly documented. Both the initial and final solutions are reported, along with intermediate solutions, if necessary.</p></td></tr><tr><td>✔</td><td><p>All preprocessing steps have been clearly documented. Abnormalities or characteristics of the data or input that needed to be addressed have been corrected. If no data preprocessing is necessary, it has been clearly justified.</p><p>You've done a really nice job walking the reader through each step in your preprocessing (including the code snippets!)</p></td></tr></table> <h3>Results</h3> <table><tr><td>✔</td><td><p>The final results are compared to the benchmark result or threshold with some type of statistical analysis. Justification is made as to whether the final model and solution is significant enough to have adequately solved the problem.</p><p>The Logistic Regressor achieved an AUC of the ROC curve of 0.81 on the training dataset (mailout_train), while the GradientBoostingClassifier had scored 0.89.</p><p>Great job beating your benchmark by quite a bit!</p></td></tr><tr><td>✔</td><td><p>The final model's qualities—such as parameters—are evaluated in detail. Some type of analysis is used to validate the robustness of the model's solution.</p><p>AUC of the ROC on Test dataset: our final GradientBoostingClassifier model obtained an AUC of the ROC curve of 0.74612 on the test dataset, which led to my current position in the Leaderboard (174th out of 270 participants, link: https://www.kaggle.com/c/udacity-arvato-identify-customers/leaderboard).</p><p>Nice job! Another approach that you could take would be to demonstrate that your tuned model is robust would be to perform a k-fold cross validation. In this case, you'd document how the model performs across each individual validation fold. If the validation performance is stable and doesn't fluctuate much, then you can argue that the model is robust against small perturbations in the training data.</p></td></tr></table> <div> DOWNLOAD PROJECT</div>			✔	<p>Student provides a high-level overview of the project in layman's terms. Background information such as the problem domain, the project origin, and related data sets or input data is given.</p> <p>Second, and most importantly, our end-goal is to become sufficiently confident in predicting, based on given demographics data, who (from the general population) would be highly probable to become a new customer of the mail-order company: these individuals would then be targeted by the mail-order company through their marketing campaign. This, in particular, would optimize their marketing action and, ultimately, their customer acquisition process, through informed and sensible customer campaign.</p> <p>Nice overview of the problem domain! I love the focus on the real-world impact of the application.</p> <p>Suggested:</p> <ul style="list-style-type: none">It's a good idea to cite some of the studies where the machine learning techniques that you're using were pioneered. This shows that you really know the field and it gives credit to the inventors.	✔	<p>The problem which needs to be solved is clearly defined. A strategy for solving the problem, including discussion of the expected solution, has been made.</p> <p>The problem can be quantified in the following terms: number of current/established customer and general population clusters (customer segmentation unsupervised problem) and probability of being a new customer to the company (supervised problem).</p> <p>You've done a great job restating the problem clearly!</p> <p>Suggested:</p> <ul style="list-style-type: none">This is a good point to begin to justify why your solution is a good 'fit' for the problem. If you were submitting this to a journal for peer review, you'd want to keep the readers focused on what you want them to think about. If they get distracted, they can ask for random things in subsequent revisions (which can significantly drag out the process and lead to arguments).	✔	<p>Metrics used to measure the performance of a model or result are clearly defined. Metrics are justified based on the characteristics of the problem.</p> <p>Our final supervised model (used to make predictions) is submitted to Kaggle, in the "Udacity+Arvato: Identify Customer Segments" competition, where the evaluation metric used is the AUC for the ROC curve, i.e. the Area Under the Curve for the Receiver Operating Characteristic (ROC) curve. The significant class imbalance in the MAILOUT_TEST dataset motivates us to use this metric, instead of, for example, a basic accuracy score (which would have heavily been impacted by the imbalance, thus not depicting an accurate picture of obtained results and accuracy).</p> <p>Excellent justification here! AUC is probably the most optimal metric for this problem. Precision and recall can also be useful for imbalanced datasets where you want to emphasize a model's performance in a specific area. MCC is another commonly used metric for imbalanced multi-class problems.</p>	✔	<p>If a dataset is present, features and calculated statistics relevant to the problem have been reported and discussed, along with a sampling of the data. In lieu of a dataset, a thorough description of the input space or input data has been made. Abnormalities or characteristics of the data or input that need to be addressed have been identified.</p> <p>Nice job documenting the scope and nature of the dataset for the reader! A sampling of the raw data is provided for the reader and you've noted several unusual properties of the dataset.</p>	✔	<p>Algorithms and techniques used in the project are thoroughly discussed and properly justified based on the characteristics of the problem.</p> <p>I think that you've given the reader a really solid overview of the methods and techniques in your project! The emphasis on the unsupervised learning algorithms is excellent.</p> <p>**Warning: a spelling error has been made in the Project Notebook file, where SVC has been incorrectly named 'Support Vector Machine' (SVM).</p> <p>You probably don't need to explicitly note this for the reader. I'd assume that anyone who was directly reviewing your code would entirely understand what you meant.</p>	✔	<p>A visualization has been provided that summarizes or extracts a relevant characteristic or feature about the dataset or input data with thorough discussion. Visual cues are clearly defined.</p> <div><p>Looks good! Your figures are all fully labeled and described in the report text.</p><p>Suggested:</p><ul style="list-style-type: none">It's a good idea to title or name your figures. This will make the report look more polished and it makes it easier to refer to your data in the text (e.g. "see Figure 1...").</div>	✔	<p>Student clearly defines a benchmark result or threshold for comparing performances of solutions obtained.</p> <p>As mentioned in the Project Proposal, we chose a basic Logistic Regression Model as a benchmark model, against which we will compare the final trained model in Part 2.</p> <p>Sounds great! This is a super common default implementation that shouldn't be too hard to beat.</p>	✔	<p>The process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented. Complications that occurred during the coding process are discussed.</p> <p>You've done a really nice job documenting your implementation for the reader. I don't think that a skilled programmer would have any difficulty reproducing your results using only the project report (since you've fully documented the model hyperparameters for each model). One other thing that I'd suggest adding here would be to note if there were any complications or difficulties that you encountered during the coding process.</p> <p>EDIT: I see you mentioned some issues you encountered with the SVC being slow to train/tune later in the report. Excellent job!</p>	✔	<p>The process of improving upon the algorithms and techniques used is clearly documented. Both the initial and final solutions are reported, along with intermediate solutions, if necessary.</p>	✔	<p>All preprocessing steps have been clearly documented. Abnormalities or characteristics of the data or input that needed to be addressed have been corrected. If no data preprocessing is necessary, it has been clearly justified.</p> <p>You've done a really nice job walking the reader through each step in your preprocessing (including the code snippets!)</p>	✔	<p>The final results are compared to the benchmark result or threshold with some type of statistical analysis. Justification is made as to whether the final model and solution is significant enough to have adequately solved the problem.</p> <p>The Logistic Regressor achieved an AUC of the ROC curve of 0.81 on the training dataset (mailout_train), while the GradientBoostingClassifier had scored 0.89.</p> <p>Great job beating your benchmark by quite a bit!</p>	✔	<p>The final model's qualities—such as parameters—are evaluated in detail. Some type of analysis is used to validate the robustness of the model's solution.</p> <p>AUC of the ROC on Test dataset: our final GradientBoostingClassifier model obtained an AUC of the ROC curve of 0.74612 on the test dataset, which led to my current position in the Leaderboard (174th out of 270 participants, link: https://www.kaggle.com/c/udacity-arvato-identify-customers/leaderboard).</p> <p>Nice job! Another approach that you could take would be to demonstrate that your tuned model is robust would be to perform a k-fold cross validation. In this case, you'd document how the model performs across each individual validation fold. If the validation performance is stable and doesn't fluctuate much, then you can argue that the model is robust against small perturbations in the training data.</p>
✔	<p>Student provides a high-level overview of the project in layman's terms. Background information such as the problem domain, the project origin, and related data sets or input data is given.</p> <p>Second, and most importantly, our end-goal is to become sufficiently confident in predicting, based on given demographics data, who (from the general population) would be highly probable to become a new customer of the mail-order company: these individuals would then be targeted by the mail-order company through their marketing campaign. This, in particular, would optimize their marketing action and, ultimately, their customer acquisition process, through informed and sensible customer campaign.</p> <p>Nice overview of the problem domain! I love the focus on the real-world impact of the application.</p> <p>Suggested:</p> <ul style="list-style-type: none">It's a good idea to cite some of the studies where the machine learning techniques that you're using were pioneered. This shows that you really know the field and it gives credit to the inventors.																									
✔	<p>The problem which needs to be solved is clearly defined. A strategy for solving the problem, including discussion of the expected solution, has been made.</p> <p>The problem can be quantified in the following terms: number of current/established customer and general population clusters (customer segmentation unsupervised problem) and probability of being a new customer to the company (supervised problem).</p> <p>You've done a great job restating the problem clearly!</p> <p>Suggested:</p> <ul style="list-style-type: none">This is a good point to begin to justify why your solution is a good 'fit' for the problem. If you were submitting this to a journal for peer review, you'd want to keep the readers focused on what you want them to think about. If they get distracted, they can ask for random things in subsequent revisions (which can significantly drag out the process and lead to arguments).																									
✔	<p>Metrics used to measure the performance of a model or result are clearly defined. Metrics are justified based on the characteristics of the problem.</p> <p>Our final supervised model (used to make predictions) is submitted to Kaggle, in the "Udacity+Arvato: Identify Customer Segments" competition, where the evaluation metric used is the AUC for the ROC curve, i.e. the Area Under the Curve for the Receiver Operating Characteristic (ROC) curve. The significant class imbalance in the MAILOUT_TEST dataset motivates us to use this metric, instead of, for example, a basic accuracy score (which would have heavily been impacted by the imbalance, thus not depicting an accurate picture of obtained results and accuracy).</p> <p>Excellent justification here! AUC is probably the most optimal metric for this problem. Precision and recall can also be useful for imbalanced datasets where you want to emphasize a model's performance in a specific area. MCC is another commonly used metric for imbalanced multi-class problems.</p>																									
✔	<p>If a dataset is present, features and calculated statistics relevant to the problem have been reported and discussed, along with a sampling of the data. In lieu of a dataset, a thorough description of the input space or input data has been made. Abnormalities or characteristics of the data or input that need to be addressed have been identified.</p> <p>Nice job documenting the scope and nature of the dataset for the reader! A sampling of the raw data is provided for the reader and you've noted several unusual properties of the dataset.</p>																									
✔	<p>Algorithms and techniques used in the project are thoroughly discussed and properly justified based on the characteristics of the problem.</p> <p>I think that you've given the reader a really solid overview of the methods and techniques in your project! The emphasis on the unsupervised learning algorithms is excellent.</p> <p>**Warning: a spelling error has been made in the Project Notebook file, where SVC has been incorrectly named 'Support Vector Machine' (SVM).</p> <p>You probably don't need to explicitly note this for the reader. I'd assume that anyone who was directly reviewing your code would entirely understand what you meant.</p>																									
✔	<p>A visualization has been provided that summarizes or extracts a relevant characteristic or feature about the dataset or input data with thorough discussion. Visual cues are clearly defined.</p> <div><p>Looks good! Your figures are all fully labeled and described in the report text.</p><p>Suggested:</p><ul style="list-style-type: none">It's a good idea to title or name your figures. This will make the report look more polished and it makes it easier to refer to your data in the text (e.g. "see Figure 1...").</div>																									
✔	<p>Student clearly defines a benchmark result or threshold for comparing performances of solutions obtained.</p> <p>As mentioned in the Project Proposal, we chose a basic Logistic Regression Model as a benchmark model, against which we will compare the final trained model in Part 2.</p> <p>Sounds great! This is a super common default implementation that shouldn't be too hard to beat.</p>																									
✔	<p>The process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented. Complications that occurred during the coding process are discussed.</p> <p>You've done a really nice job documenting your implementation for the reader. I don't think that a skilled programmer would have any difficulty reproducing your results using only the project report (since you've fully documented the model hyperparameters for each model). One other thing that I'd suggest adding here would be to note if there were any complications or difficulties that you encountered during the coding process.</p> <p>EDIT: I see you mentioned some issues you encountered with the SVC being slow to train/tune later in the report. Excellent job!</p>																									
✔	<p>The process of improving upon the algorithms and techniques used is clearly documented. Both the initial and final solutions are reported, along with intermediate solutions, if necessary.</p>																									
✔	<p>All preprocessing steps have been clearly documented. Abnormalities or characteristics of the data or input that needed to be addressed have been corrected. If no data preprocessing is necessary, it has been clearly justified.</p> <p>You've done a really nice job walking the reader through each step in your preprocessing (including the code snippets!)</p>																									
✔	<p>The final results are compared to the benchmark result or threshold with some type of statistical analysis. Justification is made as to whether the final model and solution is significant enough to have adequately solved the problem.</p> <p>The Logistic Regressor achieved an AUC of the ROC curve of 0.81 on the training dataset (mailout_train), while the GradientBoostingClassifier had scored 0.89.</p> <p>Great job beating your benchmark by quite a bit!</p>																									
✔	<p>The final model's qualities—such as parameters—are evaluated in detail. Some type of analysis is used to validate the robustness of the model's solution.</p> <p>AUC of the ROC on Test dataset: our final GradientBoostingClassifier model obtained an AUC of the ROC curve of 0.74612 on the test dataset, which led to my current position in the Leaderboard (174th out of 270 participants, link: https://www.kaggle.com/c/udacity-arvato-identify-customers/leaderboard).</p> <p>Nice job! Another approach that you could take would be to demonstrate that your tuned model is robust would be to perform a k-fold cross validation. In this case, you'd document how the model performs across each individual validation fold. If the validation performance is stable and doesn't fluctuate much, then you can argue that the model is robust against small perturbations in the training data.</p>																									

Results

✔

The final results are compared to the benchmark result or threshold with some type of statistical analysis. Justification is made as to whether the final model and solution is significant enough to have adequately solved the problem.

The Logistic Regressor achieved an AUC of the ROC curve of 0.81 on the training dataset (mailout_train), while the GradientBoostingClassifier had scored 0.89.

Great job beating your benchmark by quite a bit!

✔

The final model's qualities—such as parameters—are evaluated in detail. Some type of analysis is used to validate the robustness of the model's solution.

AUC of the ROC on Test dataset: our final GradientBoostingClassifier model obtained an AUC of the ROC curve of 0.74612 on the test dataset, which led to my current position in the Leaderboard (174th out of 270 participants, link: <https://www.kaggle.com/c/udacity-arvato-identify-customers/leaderboard>).

Nice job! Another approach that you could take would be to demonstrate that your tuned model is robust would be to perform a k-fold cross validation. In this case, you'd document how the model performs across each individual validation fold. If the validation performance is stable and doesn't fluctuate much, then you can argue that the model is robust against small perturbations in the training data.

DOWNLOAD PROJECT

RETURN TO PATH