**Machine Learning Engineer Nanodegree**

**Capstone Project**

James Gan  
May 7th, 2018

**I. Definition**

*(approx. 1-2 pages)*

**Project Overview**

In this section, look to provide a high-level overview of the project in layman’s terms. Questions to ask yourself when writing this section:

* *Has an overview of the project been provided, such as the problem domain, project origin, and related datasets or input data?*
* *Has enough background information been given so that an uninformed reader would understand the problem domain and following problem statement?*

In my job we have propensity models for customer behaviour and these are all based on using a certain machine learning technique (XGBoost). These models predict how likely it is a customer will click on certain products on the company website.

The XGBoost model was chosen as it consistently performed the highest when using a licensed data science tool (DataRobot) which compared dozens of techniques on AUC (a metric for measuring binary classification performance).

I am curious as to whether a stacking classifier method would outperform XGBoost. A stacking classifier is a type of machine learning method that uses many techniques e.g. XGBoost, random forest etc. together to produce a prediction. The main motivation for this project is a comparison of methods which will be explored in the following sections and not so much about the dataset domain. The hope is that the learnings from this project will be useful in my job as I will be able to report back on whether XGBoost can be bettered using a stacking classifier.

Some useful initial links for stacking are provided below:

* <http://blog.kaggle.com/2017/06/15/stacking-made-easy-an-introduction-to-stacknet-by-competitions-grandmaster-marios-michailidis-kazanova/>
* https://rasbt.github.io/mlxtend/user\_guide/classifier/StackingClassifier/

The project is to be based on census data from the US Census Bureau found on the UCI Machine Learning Repository website: <https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29>

Since this is similar to the datasets I encounter at work. This is typical census data you would expect to be collected and importantly has a column stating whether each person earns over $50k a year or not (this is the feature I will be predicting).

**Problem Statement**

In this section, you will want to clearly define the problem that you are trying to solve, including the strategy (outline of tasks) you will use to achieve the desired solution. You should also thoroughly discuss what the intended solution will be for this problem. Questions to ask yourself when writing this section:

* *Is the problem statement clearly defined? Will the reader understand what you are expecting to solve?*
* *Have you thoroughly discussed how you will attempt to solve the problem?*
* *Is an anticipated solution clearly defined? Will the reader understand what results you are looking for?*

The problem is to investigate whether a stacking classifier produces a higher AUC than an XGBoost model on the census dataset when predicting whether a person earns over $50k in their annual income. I am interested to see how these methods compare since XGBoost has commonly been a very high performing algorithm e.g. in Kaggle competitions.

The steps involved to solve the problem are: cleaning the dataset so it can be used to model off of, building the baseline XGBoost model, building the stacking model by experimenting with combinations of different techniques, comparing the AUC of these two models and concluding which technique is better.

The solution to the problem will be that either a stacking classifier method comprising certain techniques does produce a higher AUC on the prediction of whether somebody earns over $50k. Conversely it could be that XGBoost has a higher AUC than the stacking classifier method.

**Metrics**

In this section, you will need to clearly define the metrics or calculations you will use to measure performance of a model or result in your project. These calculations and metrics should be justified based on the characteristics of the problem and problem domain. Questions to ask yourself when writing this section:

* *Are the metrics you’ve chosen to measure the performance of your models clearly discussed and defined?*
* *Have you provided reasonable justification for the metrics chosen based on the problem and solution?*

AUC will be used as the main metric for measuring success of the models in classifying the data. This is because in relation to benefitting the analysis performed in my work we care most about the ordering/ranking of customers in terms of how likely they are to be classified as one of the two outcomes. Therefore, we are less concerned with e.g. minimizing inaccuracy of the probability that someone will be classified as one of the two outcomes.

**II. Analysis**

*(approx. 2-4 pages)*

**Data Exploration**

In this section, you will be expected to analyze the data you are using for the problem. This data can either be in the form of a dataset (or datasets), input data (or input files), or even an environment. The type of data should be thoroughly described and, if possible, have basic statistics and information presented (such as discussion of input features or defining characteristics about the input or environment). Any abnormalities or interesting qualities about the data that may need to be addressed have been identified (such as features that need to be transformed or the possibility of outliers). Questions to ask yourself when writing this section:

* *If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?*
* *If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?*
* *If a dataset is****not****present for this problem, has discussion been made about the input space or input data for your problem?*
* *Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)*

The following points were identified as interesting or needing to be handled in data pre-processing:

* The feature ‘instance weight’ should be removed. As stated in the documentation on the UCI website it was added for stratification to enable more accurate analysis but should not be used for classifier modelling.
* There are no NaN values but there are ‘Not in universe’ and ‘?’ values in a lot of columns which should be considered for imputation at the data pre-processing stage.
* The ‘Target’ feature should be changed to a binary outcome so it is easier to pass to the classifier models, i.e. 1 for >$50k and 0 for <$50k.
* ~94% of the population had <$50k for their ‘Target’ values.
* The industry codes and occupations are basically numerical representations of the major industry and occupation codes but with less detail and can be removed if a non-numerical value can be used in the models instead.
* Education can be banded into more general groupings e.g. below the 9th grade.
* The ‘year’ feature is not needed/should not be considered for this classification problem and can be removed.
* The sunbelt region is the horizontal section across south America.
* There are 3229 duplicate rows of data that could be removed.
* There are a very high number of 0s for ‘capital gains/losses’ and ‘dividends from stocks’. Greater than ~ 90% of rows have 0 values for these features.
* ~94% of ‘wages per hour’ values are also 0.
* There is a high proportion (~50%) of unknowns for major occupation and industry codes and so adds a layer of difficulty to predicting the ‘target’ feature as we don’t know the types of jobs a lot of people have.
* Unfortunately the definition of the ‘veteran benefits’ code was not found and so this column and the questionnaire for veteran’s column should be considered for dropping as it is unclear what they mean??
* Scaling for numerical variables?
* OHE?
* <https://archive.ics.uci.edu/ml/machine-learning-databases/census-income-mld/census-income.names> has the distinct values of the features

These were mostly noticed through the report generated in pandas\_profiling or basic inspection of the Excel file.

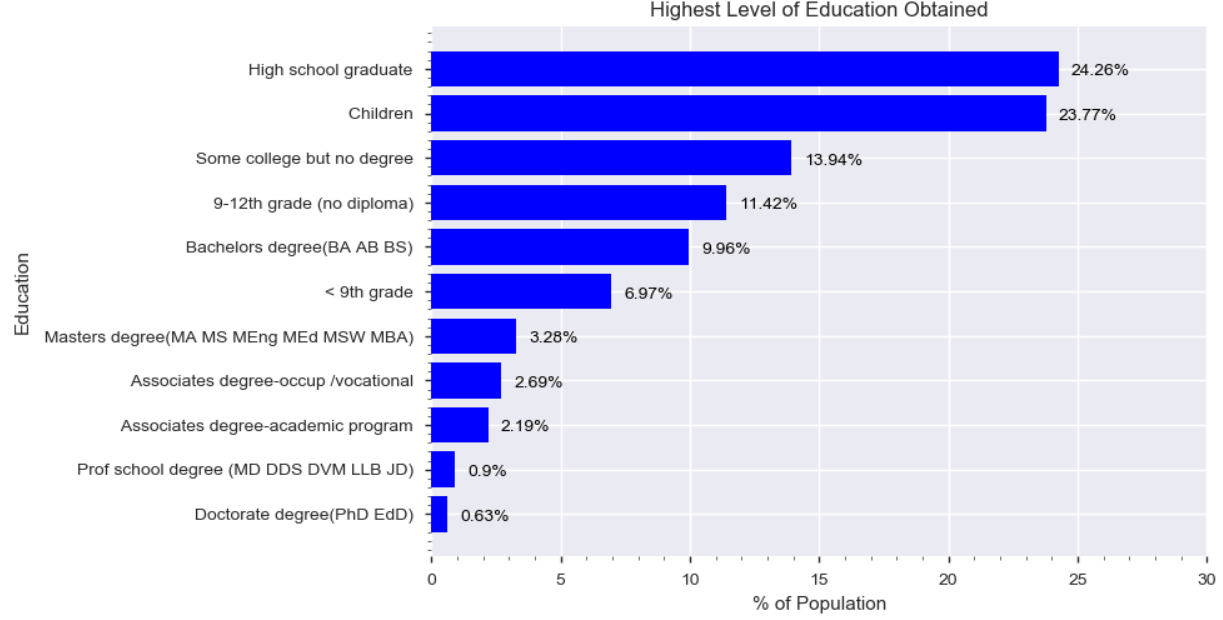
**Exploratory Visualization**

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

* *Have you visualized a relevant characteristic or feature about the dataset or input data?*
* *Is the visualization thoroughly analyzed and discussed?*
* *If a plot is provided, are the axes, title, and datum clearly defined?*

From the Data Exploration section, we have seen that there are isn’t a high proportion of data in regard to sector/job types of the survey participants. So the next best feature I have decided to analyse is level of education obtained. From looking through the Excel for the education column the ‘Children’ value is for children still in education with other values being the highest form of education received. It is also good to note that for the U.S., high school is grades 9-12.

I collapsed some of the education values into larger groupings to provide a bit more generalisation. For instance banding all those with <9th grade level education together and 9-12th grade with no diploma together.

As we can see from the chart above most people (~24%) are high school graduates and so we would not expect a lot of them to have an income >$50k. The next highest proportion of the population interestingly are children and so obviously should all have an income <$50k. These two groups make up nearly 50% of the population and so it is no surprise that there was such a high bias in the ‘Target’ feature as specified in the previous section towards <$50k earners. In fact <15% of the population has a Bachelors, Masters or Doctorate degree.

Whilst level of education is not the sole influential factor for earning power I think the two will be heavily correlated and so it is likely this will be an important feature in the models.

**Algorithms and Techniques**

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

* *Are the algorithms you will use, including any default variables/parameters in the project clearly defined?*
* *Are the techniques to be used thoroughly discussed and justified?*
* *Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?*
* XGBoost
  + Benchmark reasons
  + Random search on number of trees
* Random Forest
  + Trees and min leaf size?
  + Similar reasons to XGBoost
  + <https://stats.stackexchange.com/questions/173390/gradient-boosting-tree-vs-random-forest>
  + http://fastml.com/what-is-better-gradient-boosted-trees-or-random-forest/
* SVM
  + Large number of features?
  + Needs scaled data
  + Handles non-linear relations
* Logistic Regression
* Stacking Classifier – remove XGBoost?

**Benchmark**

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

* *Has some result or value been provided that acts as a benchmark for measuring performance?*
* *Is it clear how this result or value was obtained (whether by data or by hypothesis)?*

**III. Methodology**

*(approx. 3-5 pages)*

**Data Preprocessing**

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

* *If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?*
* *Based on the****Data Exploration****section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?*
* *If no preprocessing is needed, has it been made clear why?*

Matched up headers from name file and deleted those highlighted in red

* The feature ‘instance weight’ should be removed. more accurate analysis but should not be used for classifier modelling.
* There are no NaN values but there are, ‘Not in universe’, ‘?’ values in a lot of columns which should be considered for imputation at the data pre-processing stage.
  + ‘Not in universe or children’ ‘Not in universe under 1 year old’ not sure what these mean since these people are not children
* The ‘Target’ feature should be changed to a binary outcome so it is easier to pass to the classifier models, i.e. 1 for >$50k and 0 for <$50k.
* The industry codes and occupations are basically numerical representations of the major industry and occupation codes but with less detail and can be removed if a non-numerical value can be used in the models instead.
* Education can be banded into more general groupings e.g. below the 9th grade.
* The ‘year’ feature is not needed/should not be considered for this classification problem and can be removed.
* There are 3229 duplicate rows of data that could be removed.
* Unfortunately the definition of the ‘veteran benefits’ code was not found and so this column and the questionnaire for veteran’s column should be considered for dropping as it is unclear what they mean?? ‘own business or self employed’
* Scaling for numerical variables?
* Encoded the variables and test with/without OHE? Remove <

<https://archive.ics.uci.edu/ml/machine-learning-databases/census-income-mld/census-income.names> has the distinct values of the features

**Implementation**

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

* *Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*
* *Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*
* *Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

**Refinement**

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

* *Has an initial solution been found and clearly reported?*
* *Is the process of improvement clearly documented, such as what techniques were used?*
* *Are intermediate and final solutions clearly reported as the process is improved?*

**IV. Results**

*(approx. 2-3 pages)*

**Model Evaluation and Validation**

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

* *Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?*
* *Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?*
* *Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?*
* *Can results found from the model be trusted?*

**Justification**

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

* *Are the final results found stronger than the benchmark result reported earlier?*
* *Have you thoroughly analyzed and discussed the final solution?*
* *Is the final solution significant enough to have solved the problem?*

**V. Conclusion**

*(approx. 1-2 pages)*

**Free-Form Visualization**

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

* *Have you visualized a relevant or important quality about the problem, dataset, input data, or results?*
* *Is the visualization thoroughly analyzed and discussed?*
* *If a plot is provided, are the axes, title, and datum clearly defined?*

**Reflection**

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

* *Have you thoroughly summarized the entire process you used for this project?*
* *Were there any interesting aspects of the project?*
* *Were there any difficult aspects of the project?*
* *Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?*

**Improvement**

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

* *Are there further improvements that could be made on the algorithms or techniques you used in this project?*
* *Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?*
* *If you used your final solution as the new benchmark, do you think an even better solution exists?*