Using a Supervised Learning Model: Two-Class Boosted Decision Tree Algorithm for Income Prediction

Victor Ríos Canales
Master of Engineering in Computer Engineering
Jeffrey Duffany, Ph.D.
Electrical & Computer Engineering and Computer Science Department
Polytechnic University of Puerto Rico

Abstract *— People always try to think or ponder about things in advance regarding the decisions they make, whether if it is a good idea or even if it is not a good idea. If the person has all the correct and necessary information, good chances are that the decision will be beneficial at the end of the day. Sometimes in real life we tend to miss specific details, which need to be pondered about in order to make an assertive decision. This is when predictive analytics enters in action. These mathematical models involve statistics in order to get the most accurate and precise result so we can make an assertive decision. In these modern days where artificial intelligence is no longer a fiction movie from the past, there is a new tool called Machine Learning which we can use as a means to help us achieve precision when making or predicting an event. In this particular project the objective is to determine if there is a direct relationship between the academic educations of a person with his income.*

Key Terms *— Artificial Intelligence, Income, Machine Learning, Predictive Analytics.*

INTRODUCTION

This project consists of a new tool called Machine Learning, which is a sub field from artificial intelligence. It incorporates a wide range of methods and techniques to evaluate data and perform analysis. These analyses are made by the help of algorithms that extrapolate the data in the dataset. We will be using data from the 1994 Census, which show demographics and income. This exercise will help us better understand the relationship between education, family composition and job markets among others factors in order to predict income. The final goal is to establish if people with higher education obtain higher incomes.

HISTORY

One of the many fields in the artificial intelligence is Machine Learning. This one can be described as computing systems that improve with experience. It can also be described as a method of turning data into software. Whatever term is used, the results remain the same; data scientists have successfully developed methods of creating software "models" that are trained from huge volumes of data and then used to predict certain patterns, trends, and outcomes.

Predictive analytics is the underlying technology behind Azure Machine Learning, and it can be simply defined as a way to scientifically use the past to predict the future to help drive desired outcomes.

Machine learning and predictive analytics are typically best used under certain circumstances, as they are able to go far beyond standard rules engines or programmatic logic developed by mere mortals. Machine learning is best leveraged as means to optimize a desired output or prediction using example or past historical experiential data.
Under traditional programming models, programs and data are processed by the computer to produce a desired output, such as using programs to process data and produce a report [1].

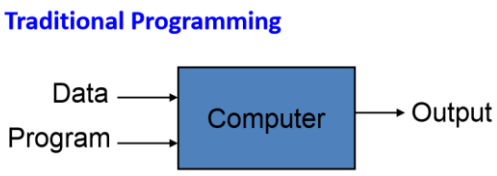

Figure 1
Traditional Programming

When working with machine learning, the processing paradigm is altered dramatically. The data and the desired output are reverse-engineered by the computer to produce a new program.

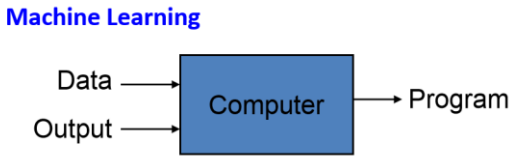

Figure 2
Machine Learning

The power of this new program is that it can effectively "predict" the output, based on the supplied input data. The primary benefit of this approach is that the resulting "program" that is developed has been trained (via massive quantities of learning data) and finely tuned (via feedback data about the desired output) and is now capable of predicting the likelihood of a desired output based on the provided data. A classic example of predictive analytics can be found everyday on Amazon.com; there, every time you search for an item, you will be presented with an upsell section on the webpage that offers you additional catalog items because "customers who bought this item also bought" those items. This is a great example of using predictive analytics and the psychology of human buying patterns to create a highly effective marketing strategy.

PURPOSE

The Two-Class Boosted Decision Tree is an algorithm seeks to predict whether a person's income exceeds $50,000 per year based on his demographics or census data.

SCHEME

The basic process of creating Machine Learning solutions is composed of a repeatable pattern of workflow steps that are designed to help create a new predictive analytics solution in no time.

- Data – It's all about the data. Here's where we acquire, compile, and analyze testing and training data sets for use in creating Azure Machine Learning predictive models.
- Create the model – Use various machine learning algorithms to create new models that are capable of making predictions based on inferences about the data sets.
- Evaluate the model – Examine the accuracy of new predictive models based on ability to predict the correct outcome, when both the input and output values are known in advance. Accuracy is measured in terms of confidence factor approaching the whole number one.
- Refine and evaluate the model – Compare, contrast, and combine alternate predictive models to find the right combinations that can consistently produce the most accurate results.
- Deploy the model – Expose the new predictive model as a scalable cloud web service, one that is easily accessible over the Internet by any web browser or mobile client.
- Test and use the model – Implement the new predictive model web service in a test or production application scenario. Adding manual or automatic feedback loops for continuous improvement of the model by capturing the appropriate details when accurate or inaccurate predictions are made [1].

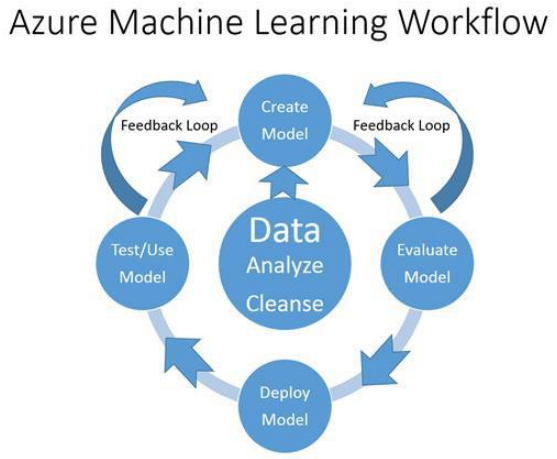

Figure 3
Machine Learning Workflow

DATABASE

The adult census income binary classification dataset would be the example of a training data set that will be used to create a new model to predict whether a person's income level would be greater or less than $50,000. This dataset and others can be located on the same ML. The data was originally gathered from the 1994 Census database and contain the following variables:

Table 1
Adult Census Income (1994)

Variables (Columns)	Description
age	Continuous.
workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt	Continuous.
education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
education-num	Continuous.
marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv,.
relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex	Female, Male.
capital-gain	Continuous.
capital-loss	Continuous.
hours-per-week	Continuous.
native-country	United States, Puerto Rico, Cambodia, England, Canada, Germany, Outlying-US (Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad & Tobago, Peru, Hong, Holand-Netherlands.
Income	<=50K, >50K.

REQUIREMENTS

Since the ML Studio run in the cloud there's no limitation of which operating system or platform you are running, but the only requirement is to have Internet access and a browser (Internet Explorer, Safari, Chrome, Firefox or any other) to access the Azure portal and the desire dataset.

TECHNOLOGIES

Microsoft Azure Machine Learning (ML) is a service that a developer can use to build predictive analytics models (using training datasets from a

variety of data sources) and then easily deploy those models for consumption as cloud web services. Azure ML Studio provides rich functionality to support many end-to-end workflow scenarios for constructing predictive models, from easy access to common data sources, rich data exploration and visualization tools, application of popular ML algorithms, and powerful model evaluation, experimentation, and web publication tooling. With ML you can build a variety of predictive analytics models using real world data, evaluate several different machines learning algorithms and modeling strategies, and then deploy the finished models as machine learning web service on Azure within a matter of minutes.

Machine Learning Algorithms

It's important to note there are several different categories of machine learning algorithms.

- Classification algorithms – These are used to classify data into different categories that can then be used to predict one or more discrete variables, based on the other attributes in the dataset.
- Regression algorithms – These are used to predict one or more continuous variables, such as profit or loss, based on other attributes in the dataset.
- Clustering algorithms – These determine natural groupings and patterns in datasets and are used to predict grouping classifications for a given variable.

In order to start learning and underlying the theories, principles, and algorithms of data science we need to understand the two main methods in which they learn about data. The first one is the supervised learning in which the prediction model is "trained" by providing known inputs and outputs. This method of training creates a function that can then predict future outputs when provided only with new inputs. Unsupervised learning, on the other hand, relies on the system to self-analyze the data and infer common patterns and structures to create a predictive model [1].

Supervised Learning

Supervised learning is a type of machine learning algorithm that uses known datasets to create a model that can then make predictions. The known data sets are called training datasets and include input data elements along with known response values. From these training datasets, supervised learning algorithms attempt to build a new model that can make predictions based on new input values along with known outcomes.

Supervised learning can be separated into two general categories of algorithms:

- Classification – This algorithm is use for predicting responses that can have just a few known values such as married, single, or divorced based on the other columns in the dataset.
- Regression – These algorithms can predict one or more continuous variables, such as profit or loss, based on other columns in the dataset.

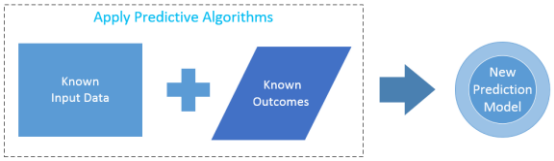

Figure 4
Supervised Learning: Apply Predictive Algorithms

Unsupervised Learning

In the case of unsupervised machine learning, the task of making predictions becomes much harder. In this scenario, the machine learning algorithms are not provided with any kind of known data inputs or known outputs to generate a new predictive model. In the case of unsupervised machine learning, the success of the new predictive model depends entirely on the ability to infer and identify patterns, structures, and relationships in the incoming data set. The goal of inferring these patterns and relationships is that the objects within a group be similar to one another and also different from other objects in other groups.

There are two basic approaches to unsupervised machine learning. One of the most common unsupervised learning algorithms is

known as cluster analysis, which is used to find hidden patterns or groupings within data sets.

Some common examples of cluster analysis classifications are the following:

- Socioeconomic tiers – Income, education, profession, age, number of children, size of city or residence, and so on.
- Psychographic data – Personal interests, lifestyle, motivation, values, involvement.
- Social network graphs – Groups of people related to you by family, friends, work, schools, professional associations, and so on.
- Purchasing patterns – Price range, type of media used, intensity of use, choice of retail outlet, fidelity, buyer or non-buyer, buying intensity.

Two-Class Boosted Decision Tree

For this project I select the Two-Class Boosted Decision Tree module to create a machine learning model that is based on the boosted decision trees algorithm. A boosted decision tree is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth. Predictions are based on the entire ensemble of trees together that makes the prediction. Generally, when properly configured, boosted decision trees are the easiest methods with which to get top performance on a wide variety of machine learning tasks. However, they are also one of the more memory intensive learners, and the current implementation holds everything in memory; therefore, a boosted decision tree model might not be able to process the very large datasets that some linear learners can handle. The construction of efficient decision trees is one of the fundamental problems in data mining. These decision trees often use similarity based metrics in order to classify newly introduced instances to one out of the predefined classes. To deal with this constraint, many heuristics have been suggested, which generally are greedy by nature, to build classification-tree models in a linear or close to linear time complexity. Most of these methods are recursive and use a top down approach: at each stage of the tree construction they often choose the best attribute with respect to some predefined optimality criterion, which evaluates the attributes "potential contribution" for a successful classification [2].

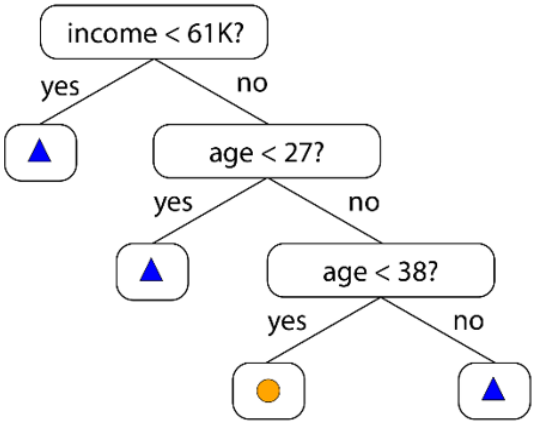

Figure 5
Decision Forest

Decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data. Data comes in records of the form: $(x,Y) = (x_1,x_2,x_3,\ldots\ldots,x_n,Y)$ when the dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector x is composed of the input variables, x1, x2, x3 etc., that are used for that task. To justify generalization, it is usually assumed that training data as well as any test data are samples from some population of (x; y) pairs [3].

Implementation

In the top its set the dataset (Adult Census Income), then the application of the Split module will help classified the data in an 80% / 20% ratio. In which 80% is going to be train or teach to evaluate the data by the Two-Class Boosted Decision Tree Model to apply the classification and the Score Model will set inferences between both before the final evaluation of each group.

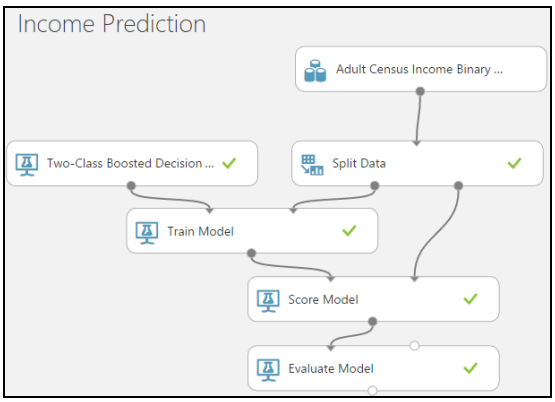

Figure 6
Income Prediction Machine Learning

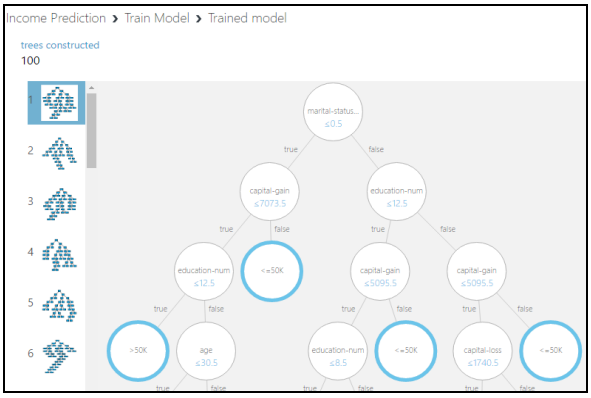

Figure 7
Top View of the First Tree

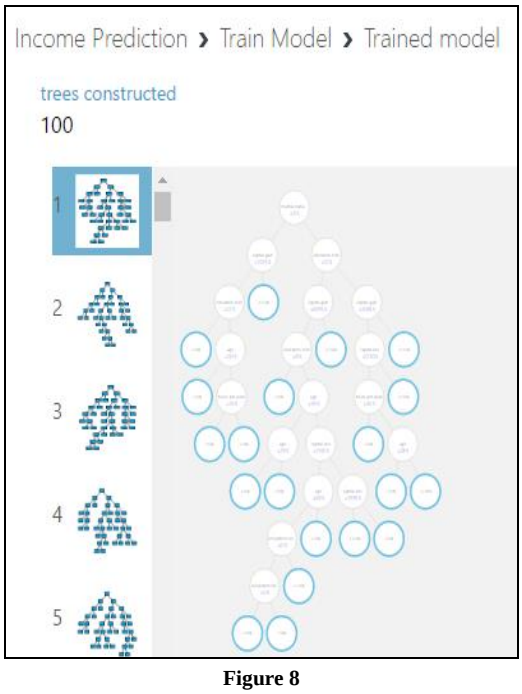

Figure 8
Complete View of the First Tree

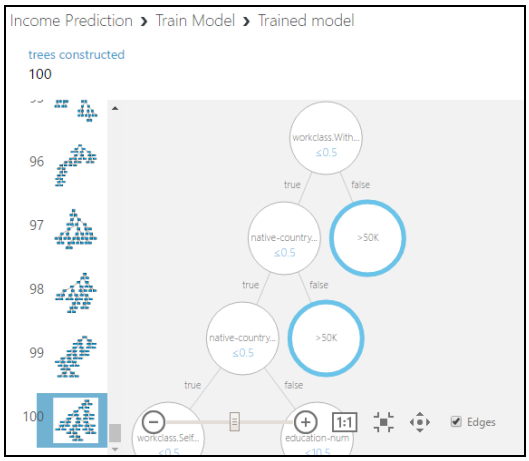

Figure 9
Top View of the Last Tree

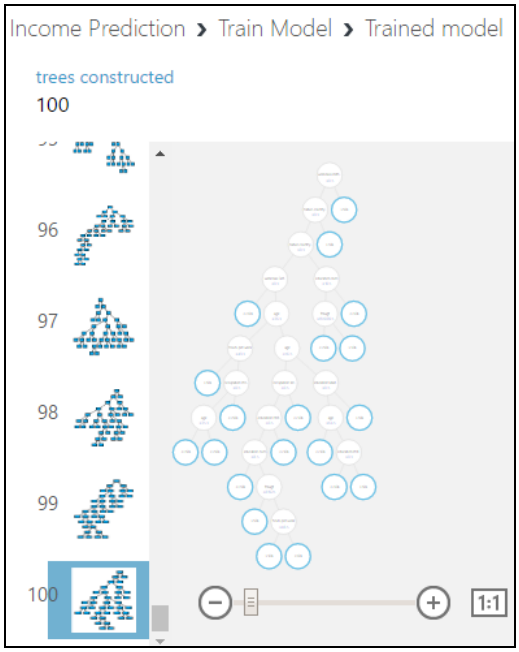

Figure 10
Complete View of the Last Tree

RESULTS

One of the biggest advantages of machine learning is that it has the unique ability to consider many more variables than a human possibly could when making scientific predictions. In this particular case we can appreciate who education plays a decisive step regarding income among different variables. Finally, when the ML is ready it can be easily deploy those models for consumption as cloud web services with an API (application program interface).

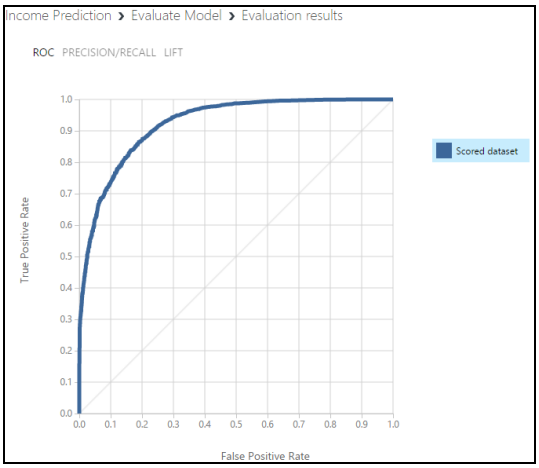

Figure 11
Evaluation Results: False Positive Rate

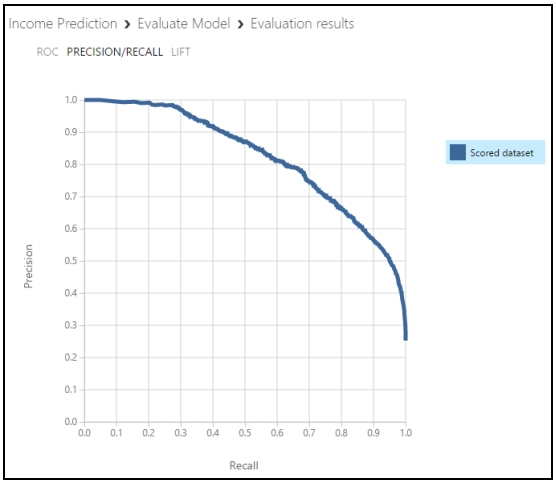

Figure 12
Evaluation Results: Precision / Recall Rate

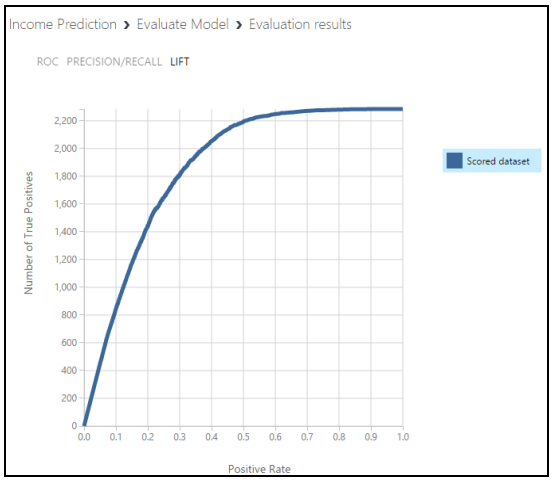

Figure 13
Evaluation Results: Positive Rate

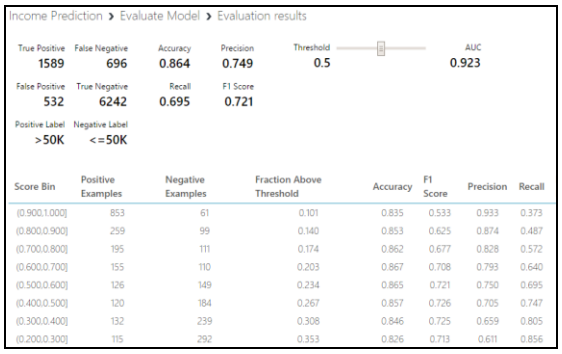

Figure 14
Final Evaluation Results

CONCLUSION

We can easily start to see patterns emerge that would likely affect the outcome based on today's common knowledge; specifically that education level and occupation are major factors in predicting the outcome. No wonder parents constantly remind their children to stay in school and get a good education. This is also the same basic process that supervised learning prediction algorithms attempt to achieve: to determine a repeatable pattern of inference that can be applied to a new set of input data.

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Jeffrey Duffany for his contribution of ideas and guidance during the project and classes. Also a special thanks to Dr. Alfredo Cruz for the guidance provided in the different courses.

REFERENCES

[1] J. Barnes, "Introduction to the science of data" in Azure Machine Learning, Redmond, WA: Microsoft Press, 2015.

[2] I. Ben-Gal et al., "Efficient Construction of Decision Trees by the Dual Information Distance Method" Department of Industrial Engineering, Tel-Aviv Univ., Israel, Quality Technology & Quantitative Management, Vol. 11, (2014) vol. 1.

[3] D. Mease et al., "Boosted Classification Trees and Class Probability/Quantile Estimation" Department of Marketing and Decision Sciences., San Jose State Univ., San Jose, CA, Journal of Machine Learning Research 8, (2007).