الجامعة الإسلامية العالمية ماليزيا

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

يونيۏرسيتي إسلام انتارابڠسا مليسيا

Kulliyyah OF Information & Communication Technology

CSC 3305 Data Science:

Project Proposal

Submitted To : Dr. Raini Hassan

Submission Date - 24 Oct 2019

Submitted by:

Group Members

1. Zian Md Afique Amin	(1631005)
2. Khan Nasik Sami	(1638153)
3. Md. Sariful Islam	(1626667)
4. Suhib Ahmad Abdulla	(1423183)

TABLE OF CONTENTS

1. Introduction ……….....………………………………………………………………..2

2. The Research Question and Hypothesis ….……………………...……………………...3

3. The Research Objectives ..……………………………………………………….......3

4. The Research Significance………………....…………………………………………….4

5. The Relevant Works/Literature Review…..…………………………………………….. .5

6. Methodology………….……………………………………………………………….7

7. Gantt Chart……………………………………………………………………………..10

8. References....…………………………………………………………………………..11

A Statistical Approach to Adult Census Income Data classification 1994

Introduction: Over the last two decades, humans have grown a lot of dependence on data and information in society and with this advent growth, technologies have evolved for their storage, analysis, and processing on a huge scale. The fields of Data Mining and Machine Learning have not only exploited them for knowledge and discovery but also to explore certain hidden patterns and concepts which led to the prediction of future events, not easy to obtain. The problem of income inequality has been of great concern in recent years. Making the poor better off does not seem to be the sole criterion to be in the quest for eradicating this issue. People of the United States believe that the advent of economic inequality is unacceptable and demands a fair share of wealth in society. This model actually aims to conduct a comprehensive analysis to highlight the key factors that are necessary for improving an individual's income. Such an analysis helps to set focus on the important areas which can significantly improve the income levels of individuals.

Our chosen problem is a prediction problem.

The Research Question and/or Hypothesis

What is income level of an individual with certain attributes:

Assessing whether an individual with certain attributes (age, education, sex, marital status and others) will earn higher or lower income levels. A higher income level is defined as an income level >50K/year while lower income level is defined as income <=50K/year. A function that predicts income level based on the fitted model and individual attribute will be provided.

What are the key features determining income level:

Identifying what are the top 5 features explaining much of the difference between low and high income levels. Determining the key features can help in policy formulation by identifying the few factors that can give most of the gains in income.

Hypothesis: Despite this limitation, using boosting algorithm the model will help in understanding how the income of a person varies depending on various factors such as the education-background, occupation, marital status, geography, age, number of working hours/week, etc.

The research objectives:

1. Import the data from the source and put it in the relevant platform. (Python/R)
2. Data Cleaning according to the needs.

3. Data Exploration, which involves analyzing each feature variable to check if the variables are significant for building the model.
4. Building a model, in which part we'll build a predictive model that will predict whether an individual earns above USD 50,000 or not.
5. Checking the accuracy of the model.
6. Load and evaluate the test data set in such a way that it does not have any null values or unnecessary predictor variables.
7. Validate the model.
8. Show how the income of a person varies depending on various factors.

The Research Significances

This prediction will help us to get usable information, which can contribute to the business domain or how the income of a person varies depending on various factors such as the educational background, occupation, marital status, geography, age, number of working hours/week, etc. And Results from the fitted classifier model show that marital status, capital gain, education, age and work hours (employment) determine much of the difference between low and high income levels.

The Relevant Works/Literature Review (a few)

No.	Year	Authors	Research problem / application	Main Techniques applied	Results	Future works (if any)
1.	2018	Sanket Biswas & Navoneel Chakrabarty	Adult Census Income Level Prediction	Gradient Boosting Classifier Model	88.16%	
2.	2014	Alina Lazar	Income Prediction via Support Vector Machine	Support Vector Machine	84.92%	Future work will identify these instances in a new separate class of unclassified instances
3	2017	Sisay Menji Bekena	Using decision tree classifier to predict income levels	Decision Tree	85%	
4.	2019	Sumit Mishra	Prediction of Income from Adult Census Income Dataset	Classification - Boosting Gradient	86.99%	
5.	2019	Bramesh S M & Puttaswamy B S	Comparative Study of Machine Learning Algorithms on CenROCsus Income Data Set	Gradient Boosting Classifier, Gaussian Naive Bayes, Support Vector Classifier,	Gradient Boosting Classifier: 90%	Make sure of having more recent census data, which will give better results on what are the best

				Decision Tree Classifier and Random Forest Classifier		**approaches for data analysis. More algorithms need to be considered.**
6.	**2017**	**Mohammed Topiwalla**	**Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting**	**Extreme Gradient Boosting (XGBOOST)**	**87.53 %**	
7.	**2018**	**C. Jayavarthini & Ishu Todi & Kshitij Kumar Agarwal**	**Analysis and Prediction of Adult Income**	**Logistic regression, decision trees, Naive Bayes, Random Forest & Gradient Boosting**	**68%**	

From this we can see that in recent years many people have tried to solve this income level prediction problem with this dataset. People tried to solve this using various methods and algorithms. Such as Gradient Boosting Classifier Model,Support Vector Machine,Decision Tree ,Gaussian Naive Bayes,Random Forest Classifier ,Extreme Gradient Boosting (XGBOOST) etc. The success rate of the prediction is varying from 85-90%.But we can see that the Gradient Boosting Classifier Model has been the most successful and consistent among them.

Methodology :

The data for our study was accessed from the University of California Irvine (UCI) Machine Learning Repository . It was actually extracted by Barry Becker using the 1994 census database. The data set includes figures on 48,842 different records and 14 attributes for 42 nations. The 14 attributes consist of 8 categorical and 6 continuous attributes containing information on age, education, nationality, marital status, relationship status, occupation, work classification, gender, race, working hours per week, capital loss and capital gain as shown in Table 1. The binomial label in the data set is the income level which predicts whether a person earns more than 50 Thousand Dollars per year or not based on the given set of attributes.

Before processing the Adult Data-set, cleaning the data with certain prepossessing techniques becomes a necessity. This includes:

1) Handling Missing Values: The data-set contains certain set of missing values for categorical features, work-class, occupation, native-country which has been dealt with some algorithmic transformations applied to the data. The missing values are flexibly handled for every attribute by setting a default marker called '?' and assigning a unique category for negating information loss.

2) Encoding of Categorical or Non-Numeric features: As all Categorical Features are non-numeric, encoding has been done in 2 stages: • Label Encoding: All categorical features are label encoded, where alphabetically each category is assigned numbers starting from 0. This is also done before running the Extra Trees Classifier Algorithm for efficient feature selection. • One-Hot Encoding: This involves splitting of different categorical features into its own

categories where each and every category assumes a binary value i.e., 0 if it does not belong to that category and 1 if it belongs to that category. This is important for those categorical features where there exists no ordinal relationship in between them. One-Hot Encoding has been done for categorical features having more than 2 categories. Here, for all categorical features except sex attribute, all label encoded forms are transformed into One-Hot Encoded Forms. This is because sex attribute has only 2 categories i.e., male and female, which have been already represented in binary form in a single attribute and hence to avoid the curse of dimension, no One-Hot Encoding is done for sex attribute.

3) Shuffling: The whole data-set has been shuffled in a consistent way such that all the categories of different attributes remain included in Training Set and Validation Set.

4) Splitting: Now, the data-set is split into training and testing sets. With 80% of the data made available for training purposes and the remaining 20% is used for testing. E.

Learning Algorithm :The learning algorithm, used to build the predictive model is an Ensemble Learning and Boosting Algorithm known as Gradient Boosting Classifier.

1) Boosting: It is an ensemble technique, in which predictors (which are decision trees) are being constructed sequentially rather than independently .

2) Implementation: In GBC, in the sequence of predictors being constructed, at each and every sequence the error in the previous decision tree is corrected by the decision tree following it. So, at each step, the GB Classifier, tends to fit the Training Data more and more.The Pseudo Code for Gradient Boosting Classifier is given below.

1. A learning rate alpha is assumed,

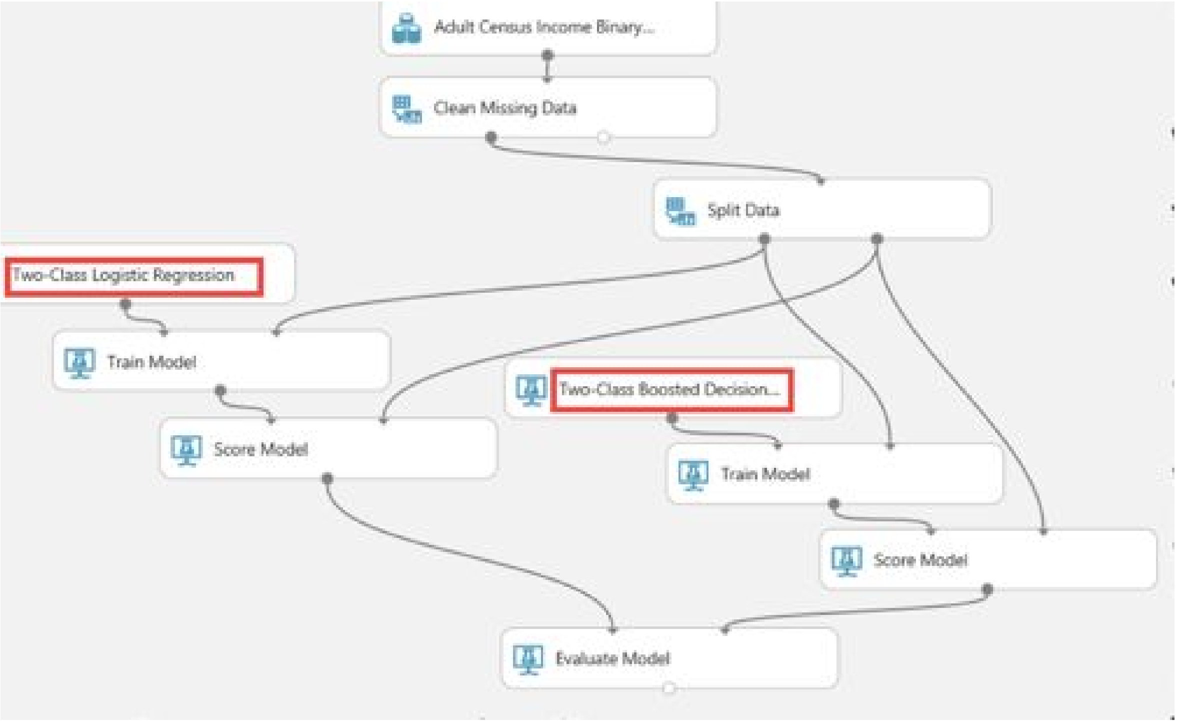
2. A weak classifier is selected .

3. The Population Distribution is updated in the next step.

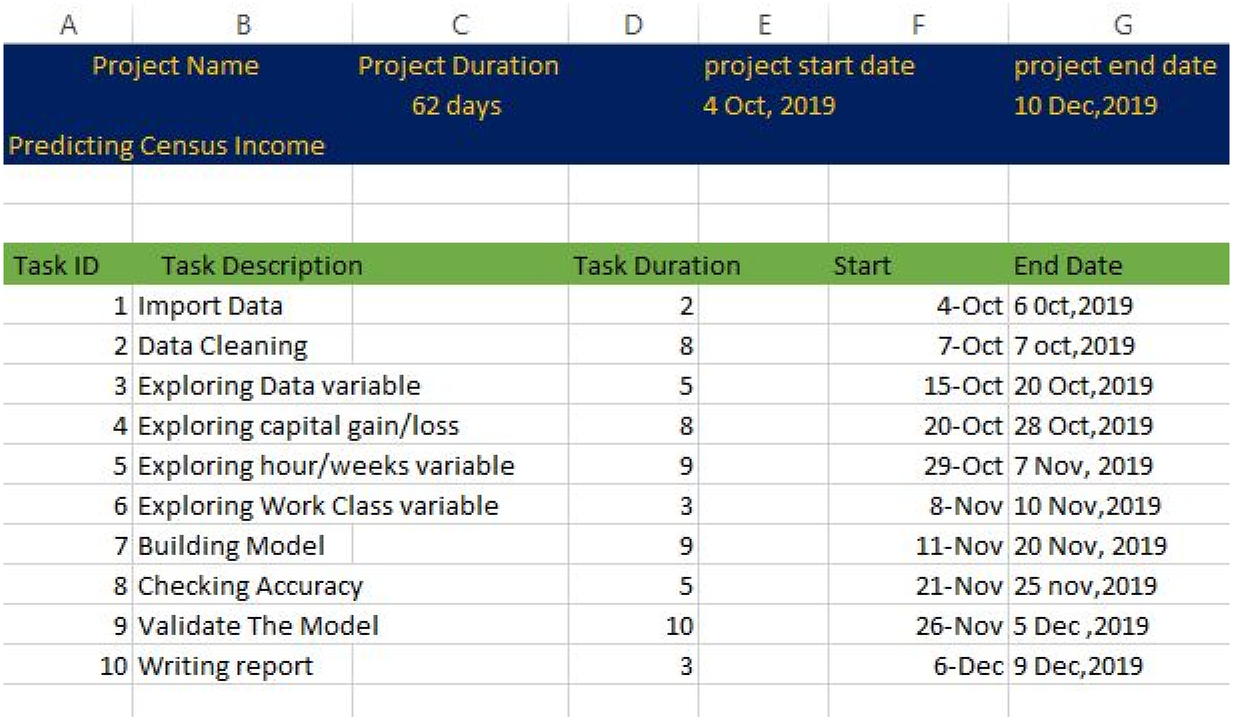
4. The new Population Distribution is used to construct the next learner or decision tree.

5. Steps 1-4 are iterated, until no hypothesis is found which can result in further improvement.

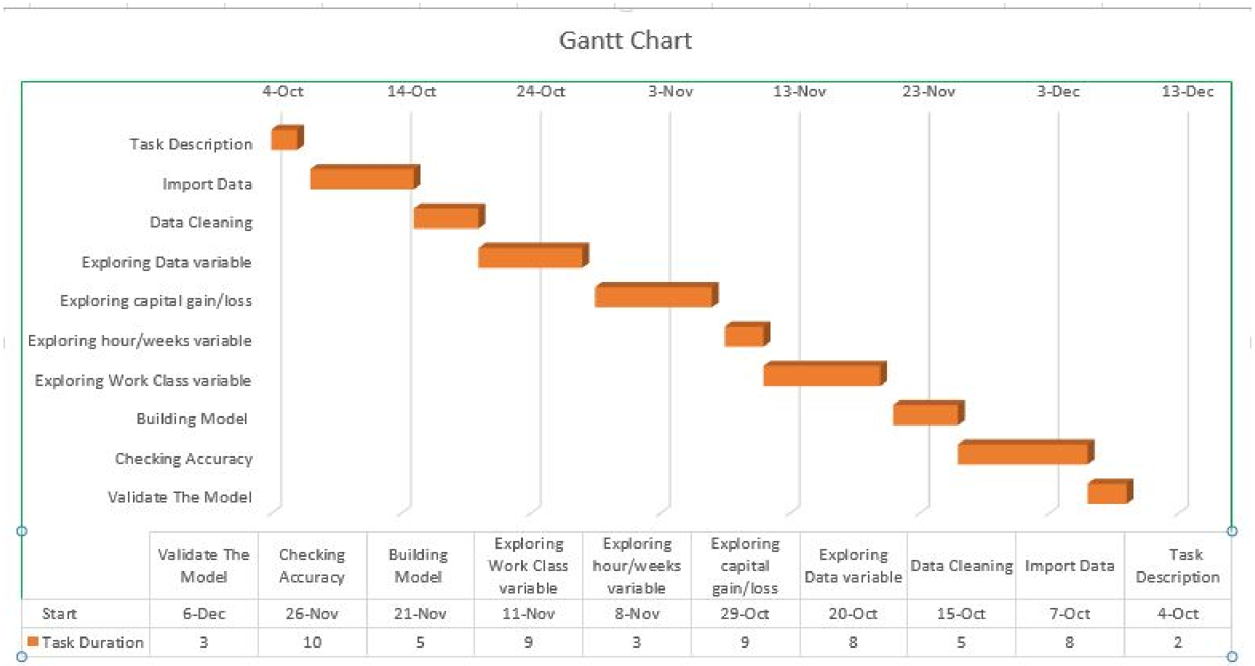
6. A weighted average of the frontier is taken using all the learners used till now where the weights are simply the alpha values.


Adult Census Income Binary...
Clean Missing Data
Split Data
Two-Class Logistic Regression
Train Model
Score Model
Two-Class Boosted Decision...
Train Model
Score Model
Evaluate Model

Gantt Chart :



A	B	C	D	E	F	G
	Project Name	Project Duration		project start date		project end date
		62 days		4 Oct, 2019		10 Dec,2019
Predicting Census Income						
Task ID	Task Description		Task Duration		Start	End Date
1	Import Data		2		4-Oct	6 0ct,2019
2	Data Cleaning		8		7-Oct	7 oct,2019
3	Exploring Data variable		5		15-Oct	20 Oct,2019
4	Exploring capital gain/loss		8		20-Oct	28 Oct,2019
5	Exploring hour/weeks variable		9		29-Oct	7 Nov, 2019
6	Exploring Work Class variable		3		8-Nov	10 Nov,2019
7	Building Model		9		11-Nov	20 Nov, 2019
8	Checking Accuracy		5		21-Nov	25 nov,2019
9	Validate The Model		10		26-Nov	5 Dec ,2019
10	Writing report		3		6-Dec	9 Dec,2019



	Validate The Model	Checking Accuracy	Building Model	Exploring Work Class variable	Exploring hour/weeks variable	Exploring capital gain/loss	Exploring Data variable	Data Cleaning	Import Data	Task Description
Start	6-Dec	26-Nov	21-Nov	11-Nov	8-Nov	29-Oct	20-Oct	15-Oct	7-Oct	4-Oct
■ Task Duration	3	10	5	9	3	9	8	5	8	2

References:

1.Sanket Biswas & Navoneel Chakrabarty (2017). Adult Census Income Level Prediction via Gradient Boosting algorithm.Jalpaiguri Government Engineering College

https://www.researchgate.net/publication/328494313_A_Statistical_Approach_to_Adult_Census_Income_Level_Prediction

2.. Alina Lazar (2014). Income Prediction via Support Vector Machine. Computer Science and Information Systems Department Youngstown State University Youngstown.

https://www.researchgate.net/publication/221226695_Income_prediction_via_support_vector_machine

3. Sisay Menji Bekena (2017). Using decision tree classifier to predict income levels.

https://mpra.ub.uni-muenchen.de/83406/1/MPRA_paper_83406.pdf

4. Sumit Mishra (2019).Classification Algorithms for the prediction of Income from Adult Census Income Dataset. Bharati Vidyapeeth's College Of Engineering, New Delhi.

https://www.academia.edu/40250231/Classification_Algorithms_for_the_prediction_of_Income_from_Adult_Census_Income_Dataset

5. Bramesh S M & Puttaswamy B S(2019). Comparative Study of Machine Learning Algorithms on Census Income Data Set. Department of Information Science & Engineering, P.E.S. College of Engineering, Mandya.

https://www.ijera.com/papers/vol9no8/Series-1/K0908017881.pdf

6.Mohammed Topiwalla(2017)Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting,University of SP Jain School of Global Management.

https://datascience52.files.wordpress.com/2017/02/machine-learning-on-uci-adult-data-set-using-various-classifier-algorithms-and-scaling-up-the-accuracy-using-extreme-gradient-boosting.pdf

7. C. Jayavarthini & Ishu Todi & Kshitij Kumar Agarwal (2018) ANALYSIS AND PREDICTION OF ADULT INCOME. Department of CSE, SRM University, Chennai, India.

https://acadpubl.eu/hub/2018-118-22/articles/22a/88.pdf