

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
|  | | Adult50K Dataset | | | | |  | |
|  |  | | | | | | |  |
|  | | | |  |  | | | |
|  | | | | Baldip Singh |  | | | |
|  | | | | 08/07/2020—Fundamentals of Business Intelligence – Quarter 2—Michael Thompson |  | | | |
|  | | |  | | |  | | |



|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | | |  |  | | | |  | | |
|  | | Scenario | | | | | | | |  | | |
|  | |  | | |  |  | | | |  | | |
|  | |  | |  | | |  | | |  | | |
|  | |  |  | Business X is a USA based group of analysts with connections to the government. Business X oversees general analysis of the population of Town X. Town X has over 30,000 inhabitants all ranging from a variety of backgrounds. Residence Information is supplied to Business X through various government databases. This data also contains a record of whether the resident earns over or less than $50,000 annual income. The government wants Business X to analyze the data in hopes of finding trends that lead to residents earning more, and thus generating more taxes. Baldip and Partner are the analysts assigned to the job, the government want Business X to focus on personal traits of individuals and whether this has any effect on the annual income. Achieving an annual income over $50,000 is the focus as the government wants to maximize income tax revenue. Baldip and Partner are asked by their manager to create different models to predict the annual income for future residents, this information will be reported back to the manager and then spread amongst Town X. This information will allow residents to set goals of achieving these personal traits with the promise of earning higher wages and eventually the government collecting more taxes. Baldip opted to model this using the Decision Tree model and Partner via Logistic Regression model. | | |  |  | |  | | |
|  | | Baldip Singh - Decision Tree Model | | | | | | |  | | |
|  | |  | | |  |  | | |  | | |
|  | Data Preparation Upon putting the data set through RapidMiner it was noticeable that some rows had missing values represented by a question mark “?”. Keeping missing values part of the data set would question the integrity of the model created. These were then decided to be removed. After using the filter function, it was noticed that only three attributes contained these question marks, Work Class, Occupation and Native Country. Filters were put in place for each of these attributes to remove any rows with a question mark.    Due to the Scenario requirement, we are required to analyze personal attributes that produce an annual income of over 50K, decision was made to remove irrelevant attributes i.e. attributes deemed not personal enough. Attributes removed were “work class”, “fnlwgt”, “occupation”, “capital gain or loss”. For the attribute’s relationship vs marital status, being similar, marital status was kept as it gave a clearer more personal understanding of the individual's immediate family background.  Similarly, for the two attributes  Education and Education Number, education number was kept. Education number was simply the number assigned to the different education levels continually increasing as education level increased i.e. 13 is bachelors 14 master 15 prof-schooling 16 doctorate. Selecting education number made for a visually pleasing decision tree as we can simply use a point on the number line as reference. For example, we can say any education number above or below a certain value, however if we were to use education names there would be 16 individual categories and no way to relate categories to one another.  These attributes were removed using the Select Attributes operator and using the Subset attribute filter, only the required attributes were left. (“Age”, “Class”, “Education-num”, “Native-Country”, “Race”, “Relationship” and “Sex”). Using the Set Role operator “Class” (annual salary), was set to the target label as this was the attribute we were trying to model. | | | | | | | | | |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | Producing the Final Model | | |  | |
|  | |  |  |  |  | |
|  | The process Modelling method selected was a decision tree. A decision tree is a diagram that shows all the different out-comes and what decisions/attributes are required to achieve these out-comes. These attributes are displayed as nodes with the highest node being the most influential on the end outcome. The number of nodes and end leaves can be altered through the different settings on RapidMiner.  Using the default settings, a decision tree was created, this tree had far too many Nodes to display. These default settings were not appropriate for our situation, this decision tree needed to have larger leaf sizes and smaller number of nodes. Further experimenting with the setting for the decision tree created a far too simple, impractical model with very few nodes. Therefore the “maximal depth” and “minimal leaf size” were altered until a model best fit for our scenario was produced. In particular “maximal depth” was decreased from 10 to 5 and “minimal leaf size” increased from 10 to 50. (Tutorials, 2020)  The criterion for the decision tree refers to the different algorithms that can be applied to the data set on which the decision tree is built upon. These options included “gain ration”, “information gain”, “accuracy”, “least square” and “Gini index”. Gini index was used as it created a model easy to understand and present in a business setting. Gini Index favours more purer partitions i.e. the nodes split at values with the greatest difference. Our target value being a binary i.e. “>50K” or “<=50K” Gini Index was viewed to be the best split visually, to display this form of information. (Gini Index, n.d.)  Following are the examples of the models that were created before producing the final decision tree. | | | | |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | The Final Model | | |  | |
|  | |  |  |  |  | |
|  | Final ModelPersonal attributes present on Model On this final decision tree, Education number, age and gender were the determining factors in achieving a salary over 50K. This is displayed by the fact that the other attributes are not present on the Decision tree. Education number is the root node and therefore is the greatest influencer of achieving our target variable class (annual salary), “ >50K” or “<=50K”. (Decision Tree, n.d.) | | | | |  |
|  | | Attribute Significance | | |  | |
|  | |  |  |  |  | |
|  | Root node - Education From this decision tree we can see that education number is the primary influencer of annual income. This means that Education of an individual has the greatest influence on whether your annual income is over 50K. Second greatest influence is the age of the individual and finally gender. (Tutorials, 2020)    Using RapidMiner, education number and class (annual salary), were chosen and a count of individuals within each education number earning >50K or <=50K were displayed. This was then migrated to excel and the percentage earning over 50K was calculated.  Education was the root node on our decision tree. Upon further evaluation we see a strong relationship between the likelihood of achieving >50K and education number. Here we see the percentage of individuals earning over 50K at each education level, Increasing and really starts to pick up almost exponentially for bachelor’s (13) and above. | | | | |  |
|  | | Attribute Significance | | |  | |
|  | |  |  |  |  | |
|  | Second node - Age Second node of our decision tree splits at age 29.5, majority of those younger than 29.5 regardless of education earned less than 50K. Using RapidMiner the select attribute, aggregate and filter operators were incorporated to find the ratio of Salary was formed within different age groups. Quite clearly it can be noted, under 30 over 90% of individuals earn less than 50K annually. It can also be noted ages 40-60 are at peak probability of earning over 50K. Second important characteristic is to be over 29.5. Third Node – Gender and Education One thing to be noted, for female’s education number appears twice. For females if we follow from the root node, we can see the requirements are more stringent for female’s vs males. Males only need to be over 12.5 education number and 29.5 years of age to have high chances of achieving over 50K salary however females require an education number of greater than 14.5.  Males over 29 and with an Education number of over 12.5 achieved 50K+ 66% of the time whereas Females required an education number of over 14.5 to achieve >50K salary 56% of the time. This is displayed by the blue and red lines at the end leaf, red being achieving >50K. There is a clear difference. Males are earning more than 50K more often than females even with lesser education. | | | | |
|  | | Attribute Significance | | |  | |
|  | |  |  |  |  | |
|  | Out of the females, only 32% earn more than 50K at education level 14(Masters) whereas males are achieving 31% likelihood of achieving more than 50K at education level 12. Also, males are double the likelihood of achieving over 50K salary at education level 14 than females.  This further supports the model. The decision tree split at genders female and male, for males it ended with a high probability of achieving over 50K annual salary, however for females there was again the criteria for education level to be above 14.5 (Masters level of education). Females need to achieve a greater education level than males, at least the level of master’s and Doctorates is required for females to maximize their chances of achieving over 50K. A male with education level of Post graduate Schooling has a higher likelihood of achieving over 50K annual income than a female with the highest level of education -PhD.  There is a clear distinction between males and females, with males having the edge. Overall, across all ages and education levels, males are earning over $50,000 annually almost 5 times more often than females. | | | | |  |
|  | | Accuracy of Model | | |  | |
|  | |  |  |  |  | |
|  | To test the accuracy of this model the “split data”, “apply model” and “performance” operators were used. Split data operator splits the data into two groups one larger than the other. Data was split 80%-20%, 80% being the set the decision tree was built on and 20% being the test. The Decision tree is modelled using the larger data set and then tested on the smaller data set using the Apply model operator, the accuracy of this model is then displayed via a confusion matrix using the performance operator. Final decision tree having an accuracy of 79.5%. Confusion Matrix Performance vector: 79.5% overall accuracy.   |  |  |  |  | | --- | --- | --- | --- | |  | true <=50K | true >50K | class precision | | pred. <=50K | 4193 | 901 | 82.31% | | pred. >50K | 338 | 601 | 64.00% | | class recall | 92.54% | 40.01% |  |   This confusion matrix displays the times the decision tree model predicted the salary correctly within the test split. For example, in the first row the class (annual salary), “<50K” was predicted correctly 4193 times and incorrectly 901 – this produces a 82% accuracy. Likewise, “>50K” class (annual salary), was predicted correctly 64% of the time. The accuracy of this specific class is low but still acceptable enough to say there is some validity of the model. (Tutorials, 2020)  Performance vector criteria was changed to accuracy to display the overall accuracy of the model. The overall accuracy was 79.5%. This number meant that roughly 8/10 times this model would be accurate, and this level of accuracy was deemed an acceptable threshold to practically apply this model. (Confusion Matrix, n.d.)  Scenario analysis was deemed a form of analysis not applicable to this data set. | | | | |
|  | | Findings for the Manager | | |  | |
|  | |  |  |  |  | |
|  | Summary In summary, being a male over the age of 30 with a minimum education of bachelor’s level is the most favorable personal trait to achieve an annual salary of over $50,000. For females, the most favorable traits are again, to be over the age of 30 but to have at least Masters level education. It is advised the town should be informed to focus highly on education at their younger years and not expect to achieve high salaries until they are at least 30 also, females will have to focus on education a lot more as they will be required to achieve higher levels of education than males. Something to think about… **Why are Education, Age and Gender relevant?**  Education, Age and Gender are shown to be of great influence on the annual income of the residents as displayed by this decision tree. It could be beneficial to do further research on to understand the underlying factors.   * Why is education a determining factor of annual salary? More skilled individuals have more critical roles within the work force and therefore may be more likely to earn more than someone with only high school education. * Why is age a determining factor of annual salary? We can assume people under 30 cannot earn over 50K annually as often due to their sheer lack of experience within the working industry. * Why does it matter if you are a female or male? The difference for males and females cannot be accounted for. I can hypothesis again the females are being underrepresented within this data set as only 1 third of this data set is females.   These are only assumptions and could have different under lying factors, but something to think about. | | | | |
|  | | References and peer evaluation | | |  | |
|  | |  |  |  |  | |
|  | References *Confusion Matrix*. (n.d.). Retrieved from WikiPedia: https://en.wikipedia.org/wiki/Confusion\_matrix  *Decision Tree*. (n.d.). Retrieved from WikiPedia: https://en.wikipedia.org/wiki/Decision\_tree  *Gini Index*. (n.d.). Retrieved from YouTube: https://www.youtube.com/watch?v=KhOP0lSrL6c&t=38s  *RapidMiner*. (n.d.). Retrieved from https://rapidminer.com/  Tutorials. (2020). RapidMiner. | | | | |