TECH-LEARNER PROFILE EVALUATION

ISM6136.001F22.92628
DATA MINING

TEAM

Lokesh Anjaneya Pothana

Rajashekar Reddy Patlori

Rajashekhar Reddy Chinthalapalli

Ajita Kandapu

1.1. Introduction

The pace at which technology is being adopted by people from remote villages to metro cities and thereby witnessing the potential of technology, equipping humans with a perspective that is never imagined. This rapid adoption of technology everywhere and watching the resulting transition inspires every individual to take up a career in the computer field. The school culture in the current era has made its curriculum so flexible that students have enough freedom to choose their electives, making them finish different concentrations per their interests. But it's expected that many students choose subjects out of confusion and end up in a state where they need clarification on their skill sets. Most of the students will need help making decisions with an abundance of opportunities available. Even though there are advisors at every stage of life to help and guide students correctly, in the current digital era, having a model in hand to make these decisions would be more helpful.

1.2. Problem Statement

Every year tens of thousands of students graduate, and most of them get stuck in their life because they need more clarity about what profile their skillset suit. In every student's life, there will be this situation where they find themselves at a crossroads and must make a decision that will lay the roadmap for their professional career in the future.

1.3. Summary

The data consists of three metrics: the number of courses completed, the number of hours spent, and the Average score secured by the students. These details were analyzed and built a model to provide the students' best suitable role. The Neural network model turned out to be the best-suited model for the experiment based on analysis. There is a high scope for implementing this in e-learning platforms and suggesting users' profiles based on subjects taken, subjects based on selected profiles, and mentor suggestions.

2. ALL ABOUT THE DATASET

The data consists of details of students from an online educational platform where the platform makers were trying to recommend a subject catalog for students using their profile

2.1 Source of Data

The dataset has been downloaded from Kaggle.com, where it is titled “Tech Students - profile prediction”
Link: https://www.kaggle.com/datasets/scarecrow2020/tech-students-profile-prediction

2.2 Data Variables

The data consists of 16 variables in total, out of which 15 are the featured variables, and 1 is the target variable

2.2.1 Featured Variables and Functionalities

Unnamed: 0 - Useless column
NAME - Name of the student
USERID - ID for each student
HOURSDATASCIENCE - Number of study hours spent on Data Science courses
HOURSBACKEND - Number of study hours spent on Web Development (Backend) courses
HOURSFRONTEND - Number of study hours spent on Web Development (Frontend) courses

NUMCOURSESBEGINNERDATASCIENCE - Number of beginner-advanced Data Science courses completed by the student
NUMCOURSESBEGINNERBACKEND - Number of beginner Web Development (Backend) courses completed by the student
NUMCOURSESBEGINNERFRONTEND - Number of beginner Web Development (Frontend) courses completed by the student
NUMCOURSESADVANCEDDATASCIENCE - Number of advanced Data Science courses completed by the student
NUMCOURSESADVANCEDBACKEND - Number of advanced Web Development (Backend) courses completed by the student
NUMCOURSESADVANCEDFRONTEND - Number of advanced Web Development (Frontend) courses completed by the student
AVGSCOREDATASCIENCE - Average score by a student in a Data Science course who has completed it
AVGSCOREBACKEND – Average score by a student in Web Development (Backend) course who has completed it
AVGSCOREFRONTEND - Average score by a student in Web Development (Frontend) course who has completed it

2.2.2 Target Variable and Functionality

PROFILE - Technology profile of the students who have done the courses

2.3 Data Preview

These are few snippets of how the data looks

Snippet 2.1

	A	B	C	D	E	F
1		NAME	USER_ID	HOURS_DATASCIENCE	HOURS_BACKEND	HOURS_FRONTEND
2	28	Stormy Muto	58283940	7	39	29
3	81	Carlos Ferro	1357218	32	0	44
4	89	Robby Constantini	63212105	45	0	59
5	138	Paul Mckenny	23239851	36	19	28

Snippet 2.2

	G	H	I	J
1	NUM_COURSES_BEGINNER_DATASCIENCE	NUM_COURSES_BEGINNER_BACKEND	NUM_COURSES_BEGINNER_FRONTEND	NUM_COURSES_ADVANCED_DATASCIENCE
2	2	4	0	2
3	2	0	0	0
4	0	5	4	0
5	0	5	7	0
6	6	11	0	4

Snippet 2.3

	K	L	M	N
1	NUM_COURSES_ADVANCED_BACKEND	NUM_COURSES_ADVANCED_FRONTEND	AVG_SCORE_DATASCIENCE	AVG_SCORE_BACKEND
2	5	0	84	74
3	5	0	67	45
4	4	1		54
5	5	3		71

Snippet 2.4

	O	P
1	AVG_SCORE_FRONTEND	PROFILE
2		beginner_front_end
3		beginner_front_end
4	47	advanced_front_end
5	89	beginner_data_science

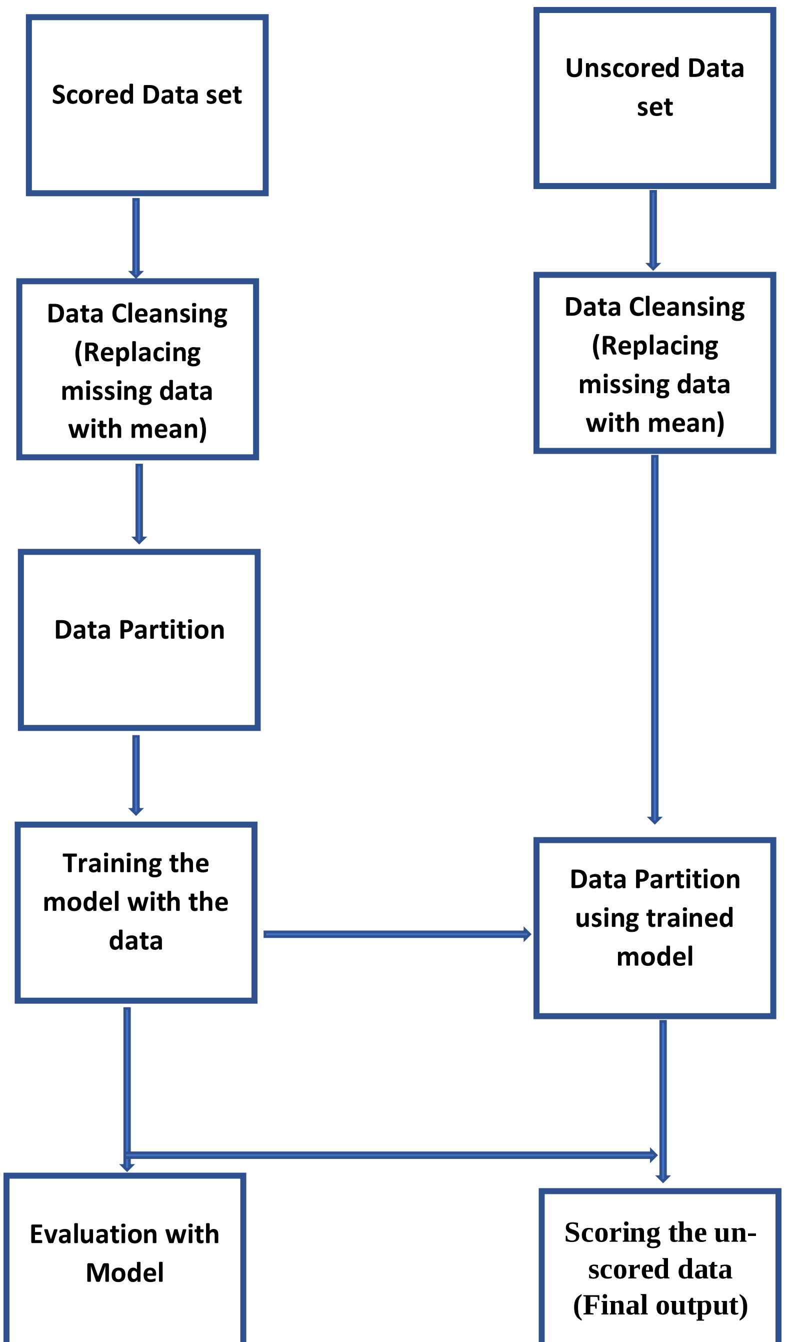
2.4 Limitations in Dataset

The data consists of the number of courses done per technology; it does not reveal what those courses are
Even though there are details of the number of hours spent on a specific technology, the data is missing elements of the number of hours spent concerning each course
The courses are divided into six categories (i.e., Adv Front End, Beginner Back End. etc.); there aren't any details of which course is designated into which technological category

3. Methodology

3.1 Flow chart of high-level overview of the complete analysis.

Figure 3.1



Multiple tools such as SAS Enterprise Miner and Azure ML Studio have been used to build in the process of developing the best model, so the data split between each model is different

3.2.1 SAS Enterprise Miner Data Splits

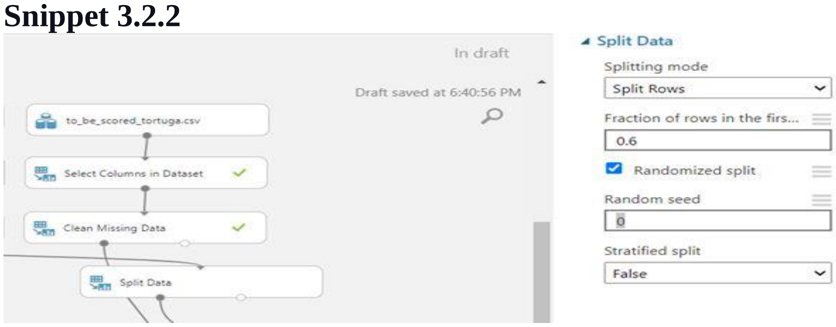
In SAS Enterprise Miner, Data Splits have been done in the ratio of 60% for training,20% for validation, and 20% for testing, and the split properties can be seen in Snippet 3.2.1

Snippet 3.2.1

3.2.2 Azure ML Studio Data Splits

In SAS Enterprise Miner, Data Splits have been done in the ratio of 60% for training and 40% for testing, and the split properties can be seen in Snippet 3.2.2

Snippet 3.2.2



3.3 Data Preprocessing

The Dataset must be Preprocessed before building as a model; there were a couple of issues that have been addressed in the Dataset

3.3.1 Deleting the insignificant columns

The Dataset has an unnamed column that has no functionality, and it has been removed from the Dataset

3.3.2 Addressing the NULL values

Some of the feature variables have NULL value records; to address the issue, the NULL values have been replaced with the median value of the column so that it doesn't affect the accuracy while building the model.

	K	L	M	N	O	P	Q	R
_COL	NUM_COL	NUM_COL	AVG_SCO	AVG_SCO	AVG_SCO	PROFILE		
2	5	0	84	74		beginner_front_end		
0	5	0	67	45		beginner_front_end		
0	4	1		54	47	advanced_front_end		
0	5	3		71	89	beginner_data_science		
4	3	0	66	85		advanced_front_end		
4	5	0	66	75		advanced_front_end		

3.4 Methods used.

3.4.1 Classification

Different classification methods were used to observe the accuracies that each model gives. The models used are 1) Decision Tree 2) Gradient Boosting 3) MBR 4) Ensemble 5) Multiclass Decision Forest 6) Multiclass Decision Jungle 7) Multiclass Logistic Regression 8) Multiclass Neural Network 9) Two-Class Boosted Decision Tree 10) Two-Class Decision Forest 11) Two-Class Decision Jungle 12) Two-Class Logistic Regression 13) Two-Class Neural Network. Azure ML and SAS Enterprise Miner tools were used to build the experiments using these models.

3.4.2 Linear regression.

It is important to know the significance in data that is used in the experiment to make more sense of the results. R studio is used to run the linear regression on the scored data set. The following results are observed.

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.5326 -1.2422  0.0264  1.1357  5.0753
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   0.4336864  0.0594364   7.297 3.06e-13 ***
hours_datascience            -0.0017199  0.0004830  -3.561 0.000371 ***
hours_backend                 0.0141141  0.0004953  28.495  < 2e-16 ***
hours_frontend                0.0020592  0.0005183   3.973 7.11e-05 ***
num_courses_beginner_datascience  0.2250445  0.0058053  38.765  < 2e-16 ***
num_courses_beginner_backend    -0.1113901  0.0051136 -21.783  < 2e-16 ***
num_courses_beginner_frontend    0.1692626  0.0048694  34.761  < 2e-16 ***
num_courses_advanced_datascience  0.0752512  0.0051030  14.746  < 2e-16 ***
num_courses_advanced_backend     0.0044839  0.0050138   0.894 0.371158
num_courses_advanced_frontend    0.0182421  0.0055242   3.302 0.000961 ***
Residual standard error: 1.48 on 19990 degrees of freedom
Multiple R-squared:  0.2497,	Adjusted R-squared:  0.2494
```

F-statistic: 739.2 on 9 and 19990 DF, p-value: < 2.2e-16

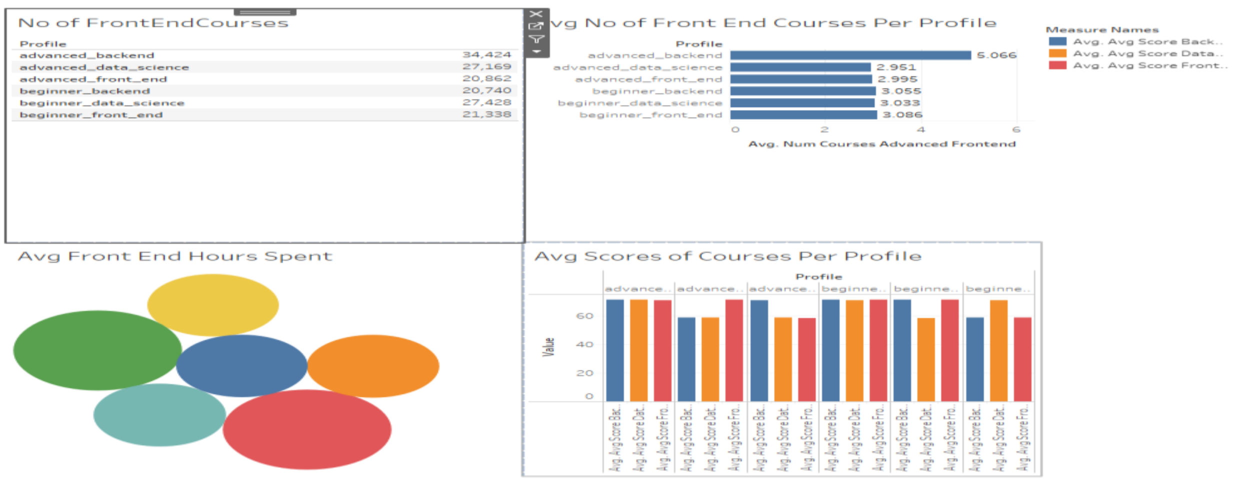
The p-vale is much less than 0.05 and Multiple R-squared is very less which implies that data is very good. Almost all the featured variables have p-value much less than 0.05 which means the data is significant to find target variable using featured variables.

3.4.3 Neural networking

In the experiment Multiclass Neural network gave highest accuracy with 95%. The main requirement in the project is to get good accuracy. So Neural network model is best fit for the project. The parameters used in this model are of single parameter, hidden layers are fully connected case, Number of hidden nodes are 20, the learning rate is 0.1, Number of learning iterations are 100 and the initial learning weights diameter is 0.1.

3.4.4 Data visualization

To understand and get more insights of the data visualization is used using Tableau



4. Related Work

The data and problem was taken from the kaggle there are three different authors EDA , FE and multiclass classification using python, UTS Probabilities Statistics using R and correlation analysis using R. The authors names and details are not available.

The following link have all the details regarding data and problem.

https://www.kaggle.com/datasets/scarecrow2020/tech-students-profile-prediction

5. 1 Results

Metrics

Overall accuracy	0.9225
Average accuracy	0.974167
Micro-averaged precision	0.9225
Macro-averaged precision	0.923491
Micro-averaged recall	0.9225
Macro-averaged recall	0.922353

Predicted Class

Actual Class	advanced...	advanced...	advanced...	beginner...	beginner...	beginner...
advanced...	92.8%	2.4%	1.8%	1.5%	0.7%	0.9%
advanced...	0.9%	94.2%	1.1%	1.0%	1.9%	0.9%
advanced...	1.2%	2.4%	90.6%	1.9%	2.9%	1.0%
beginner...	1.9%	1.9%	1.8%	89.2%	4.0%	1.2%
beginner...	1.0%	1.5%	0.7%	1.2%	94.6%	1.0%
beginner...	0.5%	2.2%	1.3%	1.5%	2.5%	92.1%

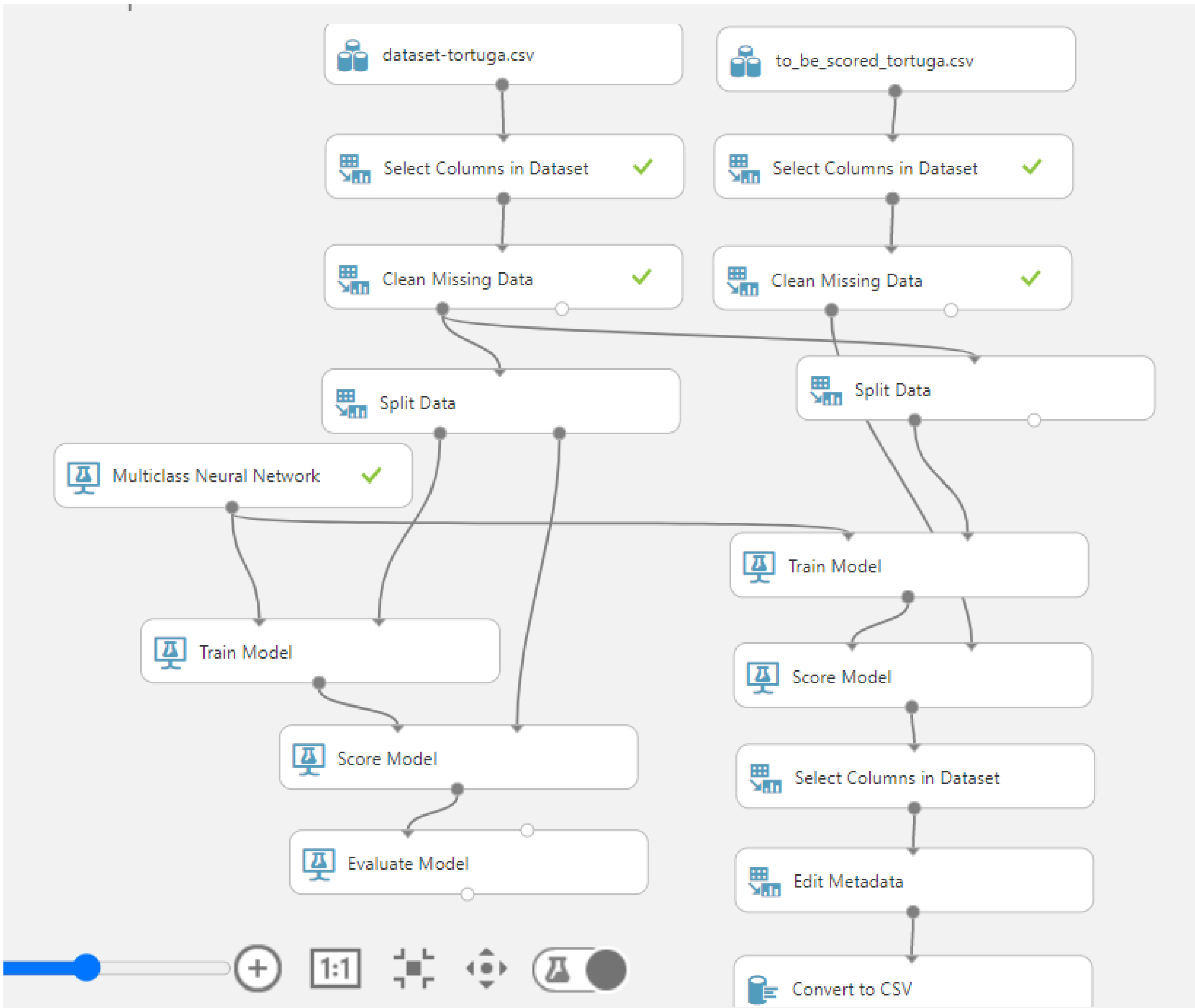
Different models are used which has been specified above and have noticed that high accuracy models is multiclass neural network with 92%. And have chosen this multiclass neural network because the model must predict the target that has multiple values. And from the confusion matrix it is clear that the model is predicting advanced_front_end 92.8% correctly, advanced backend end 94.2% accurately and advanced_data_science 90.6% perfectly. Along with this it is finding beginner_ front_end 89.2% truly, beginner_ back_end 94.6% exactly and beginner_data_science 92.1% precisely.

5.2 Conclusion

Depending on the analysis conveyed, it can be concluded that the model with multiple class neural network performs more accurately in finding the proper role for the individuals based on their course profile. Future exploration can be done by collecting alumni data and testing the model's efficiency. E-learning applications can use the model to provide suggestions. If any folks provide the details of their coursework on the website, the model can suggest the best profile fit for them and provide specific roles in which they are interested, and the model can recommend the best possible coursework to achieve their goal

6. Appendix

Snippet 6.1 AzureML workflow diagram



Snippet 6.2 - SAS Enterprise miner workflow diagram