**Virtual Learning Environment Result Prediction**

**Big Data Analytics (INSY 5376)**

**Team Members**

Asif Shaikh

Chaitanya Gokhale

Arun Tom

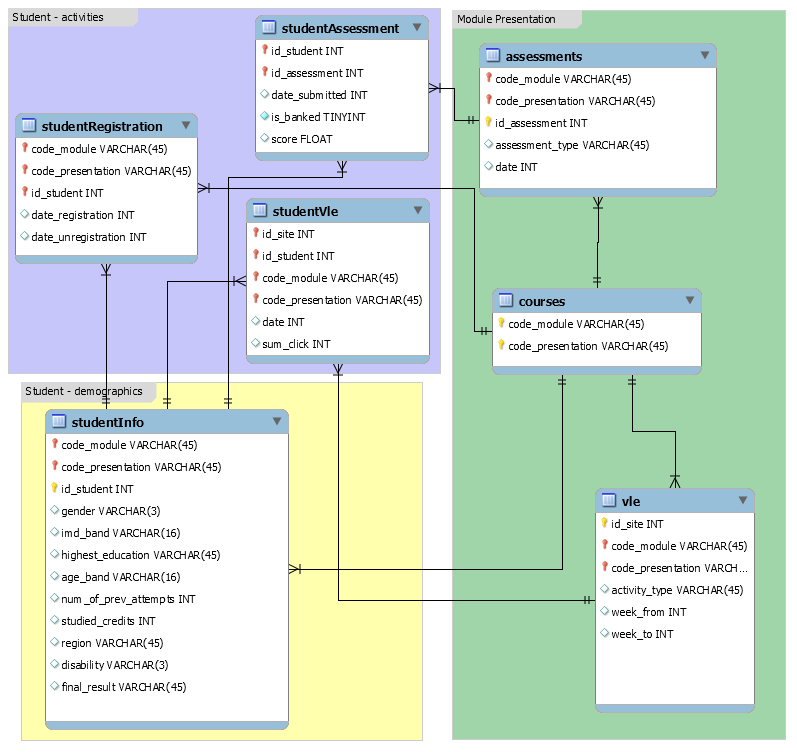
Arbaaz Shaikh

**Problem Statement**

Predict whether the student enrolled will pass or fail the Course Assessment using Machine Learning Algorithms. Take Insights from the data using Data Visualization Software Tableau.

**About the Data Set**

Dataset is taken from <https://analyse.kmi.open.ac.uk/open_dataset#data>. It is an Open University Learning Analytics Dataset (OULAD). It contains data about courses, students and their interactions with Virtual Learning Environment (VLE) for seven selected courses (called modules). Presentations of courses start in February and October - they are marked by “B” and “J” respectively. The dataset consists of 7 tables connected using unique identifiers. All tables are stored in the csv format. The schema of the given csv files is given below.



**Data Set Description**

**courses.csv**

This File contains the list of all available modules and their presentations. The Attributes are

* code\_module
* code\_presentation
* length

### assessments.csv

This file contains information about assessments in module-presentations. Usually, every presentation has number of assessments followed by the final exam. It has following Attributes.

* code\_module
* code\_presentation
* id\_assessment
* assessment\_type
* date
* weight

If the information about the final exam date is missing, it is at the end of the last presentation week.

### vle.csv

The csv file contains information about the available materials in the VLE. Typically these are html pages, pdf files, etc. Students have access to these materials online and their interactions with the materials are recorded. The vle.csv file contains the following Attributes.

* id\_site
* code\_module
* code\_presentation
* activity\_type
* week\_from
* week\_to

### studentRegistration.csv

This file contains information about the time when the student registered for the module presentation. For students who unregistered the date of unregistration is also recorded. File contains five Attributes.

* code\_module
* code\_presentation
* id\_student
* date\_registration
* date\_unregistration

Students, who completed the course have this field empty. Students who unregistered have Withdrawal as the value of the final\_result in the studentInfo.csv file.

### studentAssessment.csv

This file contains the results of students’ assessments. If the student does not submit the assessment, no result is recorded. The final exam submissions is missing, if the result of the assessments is not stored in the system. This file contains the following Attributes.

* id\_assessment
* id\_student
* date\_submitted is\_banked
* score

**studentVle.csv**

The studentVle.csv file contains information about each student’s interactions with the materials in the VLE. This file contains the following Attributes.

* code\_module
* code\_presentation
* id\_student
* id\_site
* date
* sum\_click

**Final Attributes** ( All tables are joined to create Final Data Set)

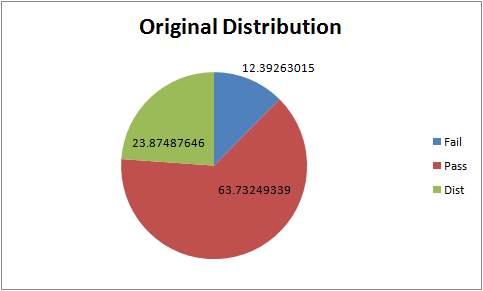
1. sum\_click– Number of times a student interacts with the material in that day.
2. date\_registration – the date of student’s registration on the module presentation, this is the number of days measured relative to the start of the module-presentation (e.g. the negative value -30 means that the student registered to module presentation 30 days before it started).
3. date\_unregistration – date of student unregistration from the module presentation, this is the number of days measured relative to the start of the module-presentation. Students, who completed the course have this field empty. Students who unregistered have Withdrawal as the value of the final\_result column in the studentInfo.csv file.
4. code\_module - an identification code for a module.
5. code\_presentation - the identification code of the presentation
6. id\_student - a unique identification number for the student.
7. gender - the student’s gender
8. region - identifies the geographic region, where the student lived while taking the module-presentation.
9. highest\_education - highest student education level on entry to the module presentation.
10. Imd\_band - specifies the Index of Multiple Depravation band of the place where the student lived during the module-presentation.
11. age\_band - band of the student’s age.
12. num\_of\_prev\_attempts - the number times the student has attempted this module.
13. studied\_credits-  the total number of credits for the modules the student is currently studying.
14. Disability -  indicates whether the student has declared a disability.
15. id\_assessment – the identification number of the assessment
16. date\_submitted - the date of student submission, measured as the number of days since the start of the module presentation.
17. is\_banked - a status flag indicating that the assessment result has been transferred from a previous presentation.
18. final\_result- Target Variable (final\_result – student’s result in the module-presentation)

**Values of Target Variable (final\_result )**

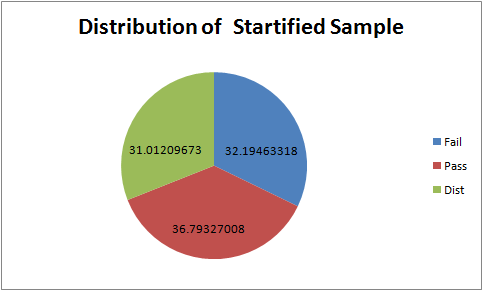
1. Pass
2. Fail
3. Distinction
4. Withdrawn(Dropped) : we dropped records taking this value as a student can withdraw from a course for various personal reasons that cannot be accounted for in a VLE environment.

**Data Preprocessing**

The total number of records in our dataset are about 200k. But the data was highly skewed in favor of pass value. The original distribution is shown below.



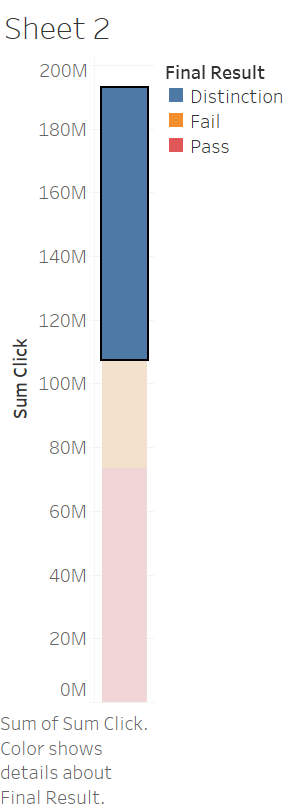
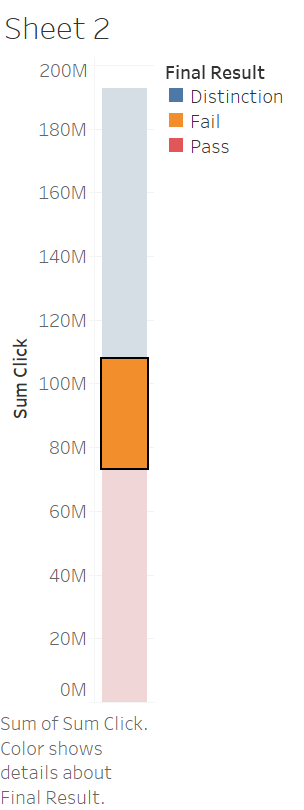
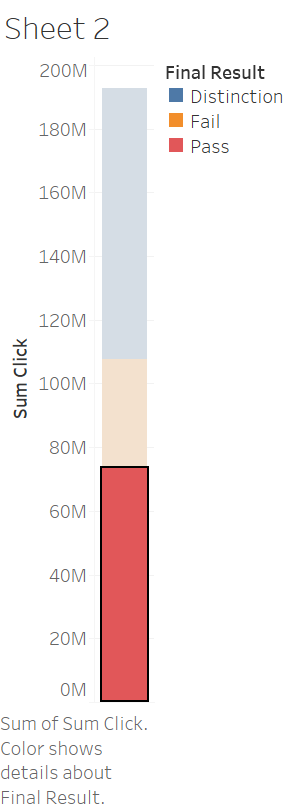
We used Random Stratified Sampling method to pick out equal samples of pass, distinction and fail values. Categorical Values are replaced with numerical indexes using stringIndexer. The Stratified Sample distribution is shown below.



**Insights from the Data**

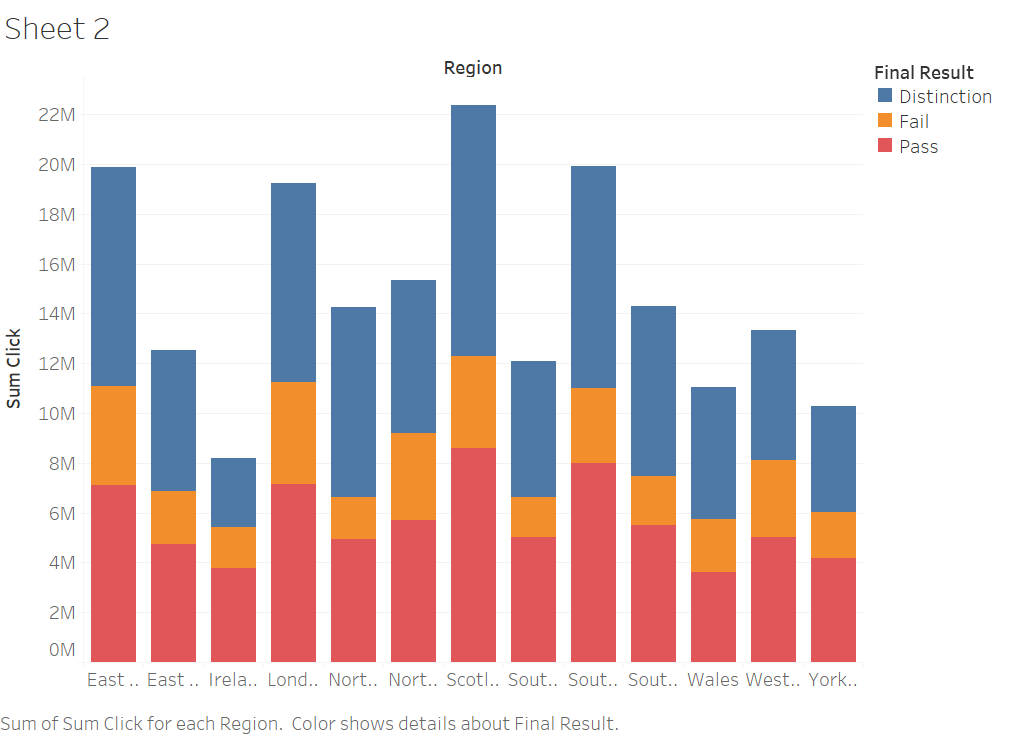
* 1. **Correlation between Sum of Clicks and Final Result:**

Sum of Clicks of Students who passed the assessment is higher than that of the students who failed. Sum of Clicks shows how much students interacted with the course over a course duration. Important insight taken is the students who are failing the course are not doing enough efforts compared to the students who are passing the course.



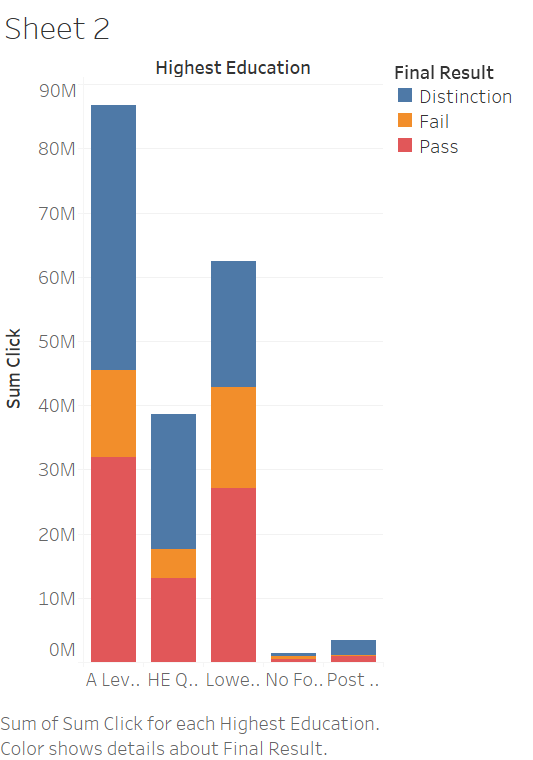
* 1. **Region based Data**

Students from Scotland has a higher participation rate in the courses while students from Ireland has the lowest participation rate. Pass, Fail and Distinction ratio is mostly constant across the region.



**3. Education based distribution:**

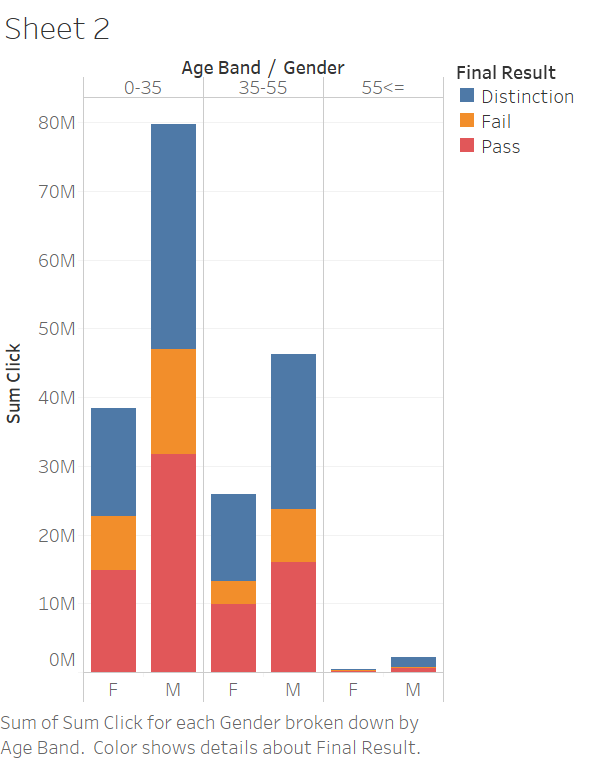
There are more participants who have completed A- level Education. Sum of Clicks Class distribution amongst the students who have Lowest level of formal education is nearly same.



**4. Age and Gender Based Distribution**

There is a large variation in Sum of Clicks between Males and Females even though there is not much difference in male and female population. It is showing that on average Males are interacting more with the material than Females.

|  |  |
| --- | --- |
| **Male** | **55000** |
| **Female** | **41000** |



**Prediction (Machine Learning Models)**

We divided the data into training set and testing set with 70% of data in training set and remaining in the test set. Machine Learning Algorithms Random Forest, Decision Tree and Support Vector Machines are trained using training set and the ML models built on training set are used to predict Result. Out of the 3 Algorithms, Random forest algorithm shown highest accuracy in predicting the result of student.

**Machine Learning Algorithm Results**

|  |  |
| --- | --- |
| Random Forest | Decision Tree |
| **F-1 Score**:  0.877867002572 | **F-1 Score:** 0.843524338927 |
| **Confusion Matrix**  [[ 5508.   466.   132.]  [  558.  5729.   746.]  [   91.   334.  5489.]] | **Confusion Matrix**  [[ 5368.   516.   254.]  [  822.  5531.   861.]   [  143.   416.  5338.] |

**Conclusion**

We have predicted the Students Final Result with 87% accuracy using Random Forest Algorithm. Good Insights are taken by visualizing the Data in Tableau.

**Potential**

If the model predicts a good grade for a student, advanced courses can be suggested for him/her to improve his skills, if negative result is predicted for him/her, we could provide supplemental material or aid depending on the interaction between student and VLE to motivate the student.

As the relation between student and the suggested material or advanced courses improve we can sell more courses/materials.