

**FINAL PROJECT: CRISP DM**

**PREDICTING STUDENT GRADES**

**BY TEAM QUAD SQUAD**

**BRIAN BUMPERS**

**AMER IMSIC**

**ASHRITH MADAN**

**SONIKA SHIVANI VIJAYKUMAR**

**TABLE OF CONTENTS**

1. **Data Collection Report**
2. **Data Description Report**
3. **Data Exploration Report**
4. **Data Quality Report**
5. **Data Preparation Report**
6. **Data Modeling Report**
7. **Results and Recommendations**
8. **References**

# **Data Collection Report:**

The data for this project was acquired from [www.kaggle.com](http://www.kaggle.com)

The link to data is <https://www.kaggle.com/daviddraper1518/predicting-student-grades> , which is a public domain database.

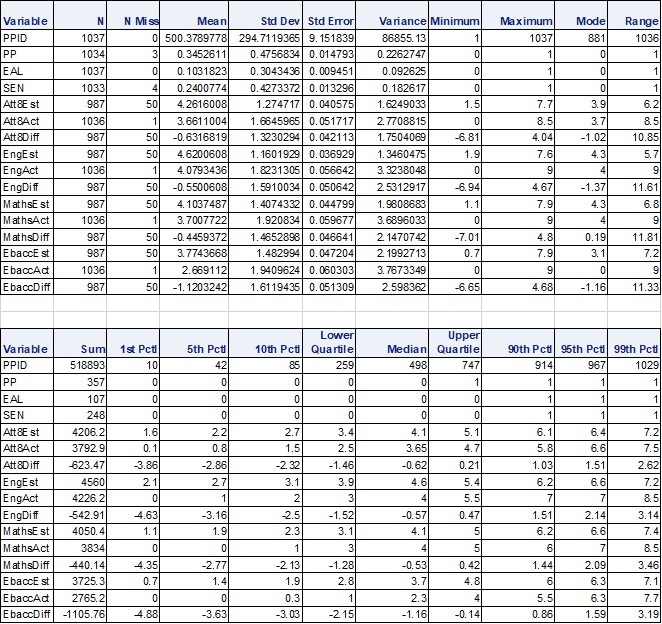
No other data sources were used for this project.

Context:  
These are the previous GCSE examination grades for the last four years at an English Underperforming School. The idea is to find a ML model that would predict future grades for students based on their profile (SEN,PP,HML,Reading, Writing and Maths Grades).

Idea is to predict the following grades (1-9 whole grades) Att8Act = Attainment 8 Actual (Best 8 grades for students) EngAct = English grade MatAct = Maths grade

PP = Pupil Premium EAL = English as Additional Language HML = Ability based on previous results ages 10/11. H= Higher, M = Middle, L = Lower Re = Reading grade at aged 10 Wr = Writing grade at aged 10 Ma = Maths grade at aged 10

# **Data Description Report:**



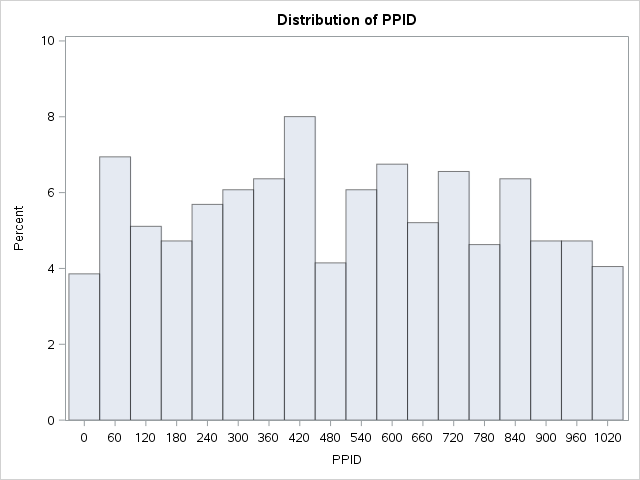
Correlation Matrix:

# 

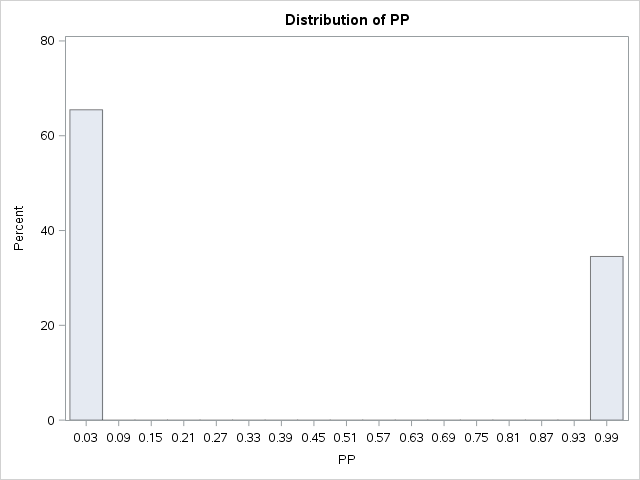
# **Data Exploration Report**

During this process, we detailed the relationships between each variable/feature with the dependent variable/label. The histograms were found as below:

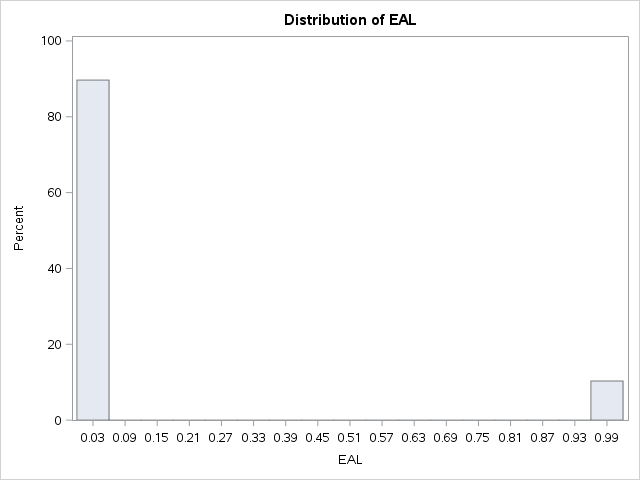
1. **PPID**



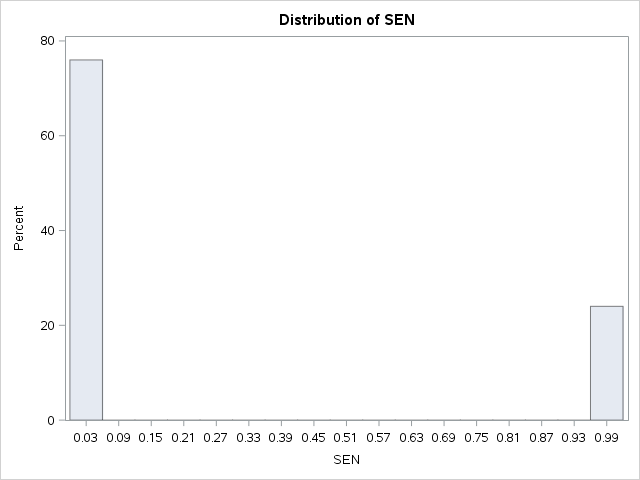
1. **PP**



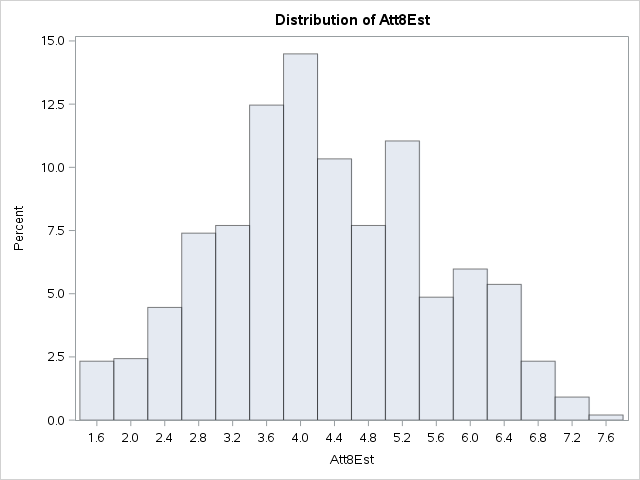
1. **EAL**



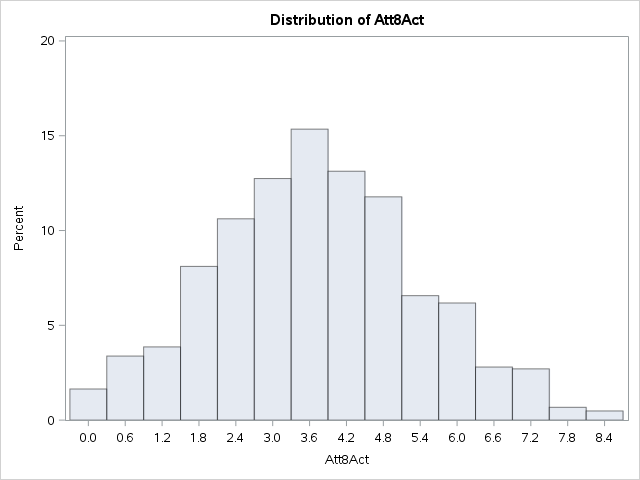
1. **SEN**



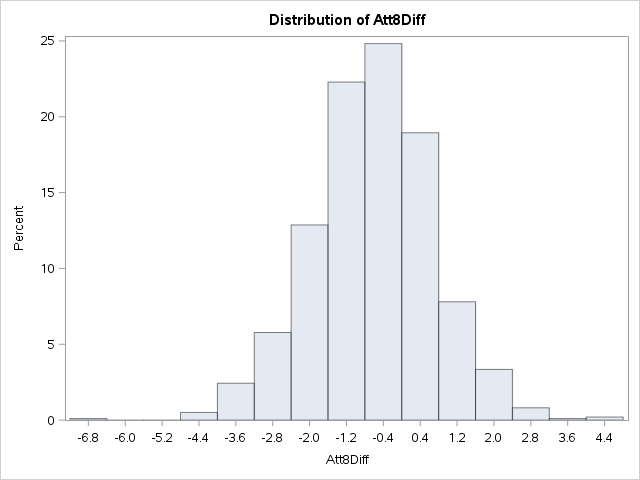
1. **Att8Est**



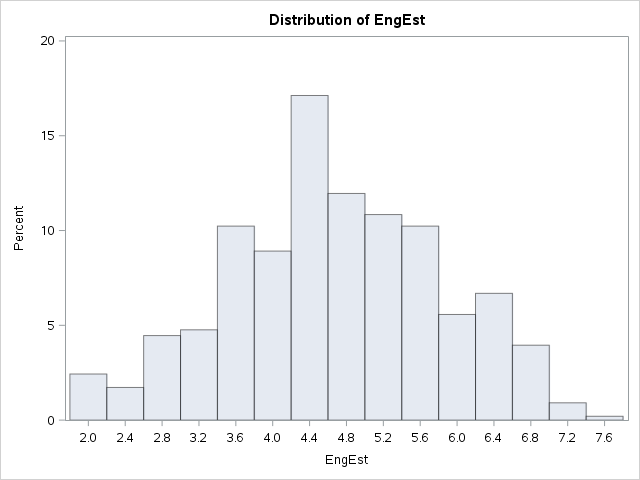
1. **Att8Act**



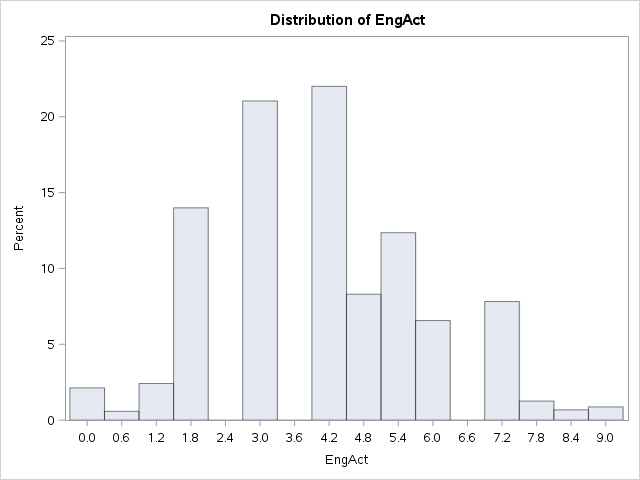
1. **Att8Diff**



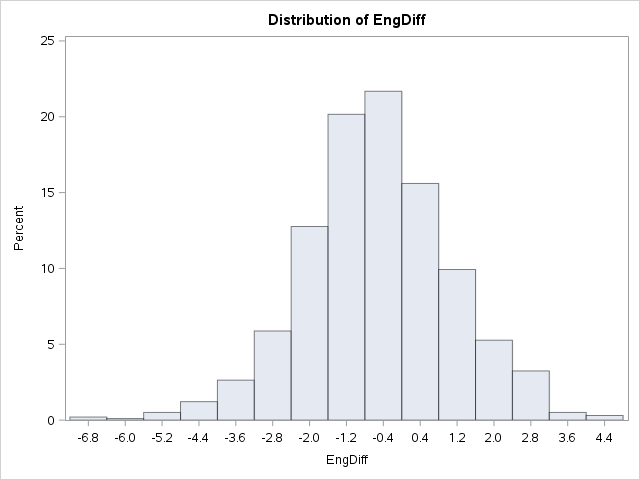
1. **EngEst**



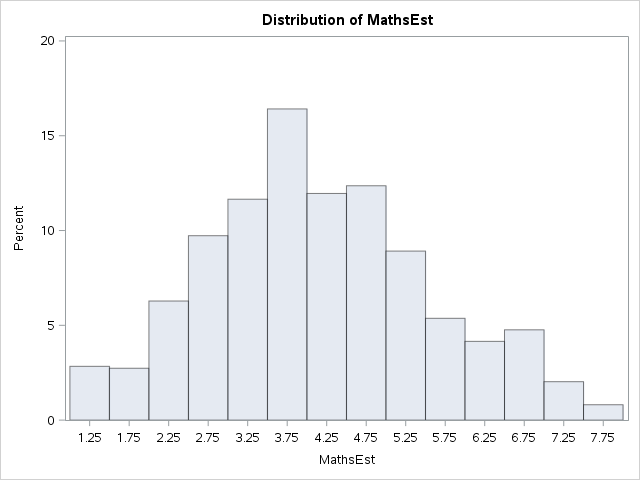
1. **EngAct**



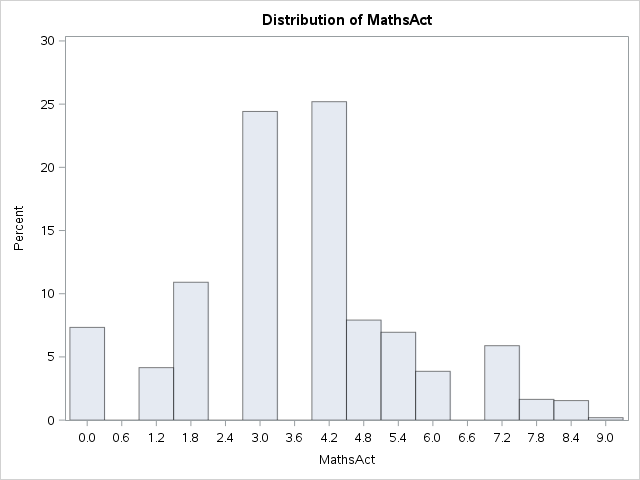
1. **EngDiff**



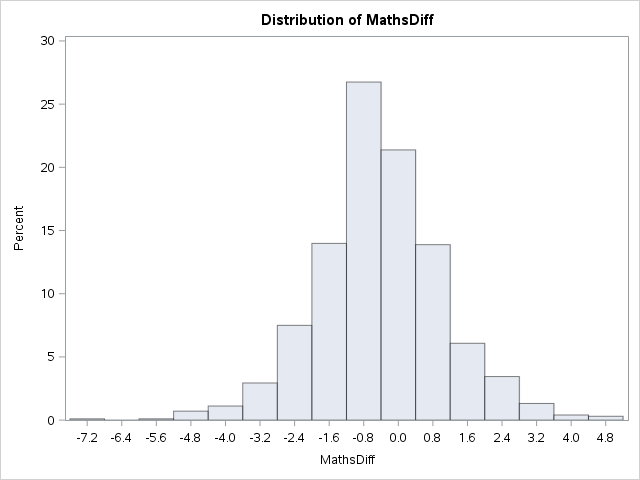
1. **MathsEst**



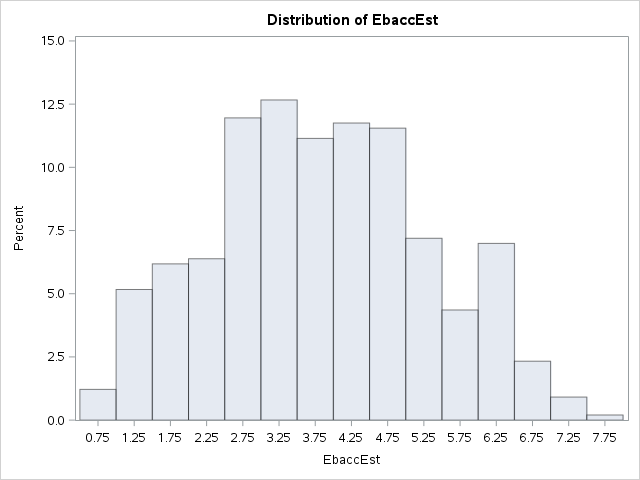
1. **MathsAct**



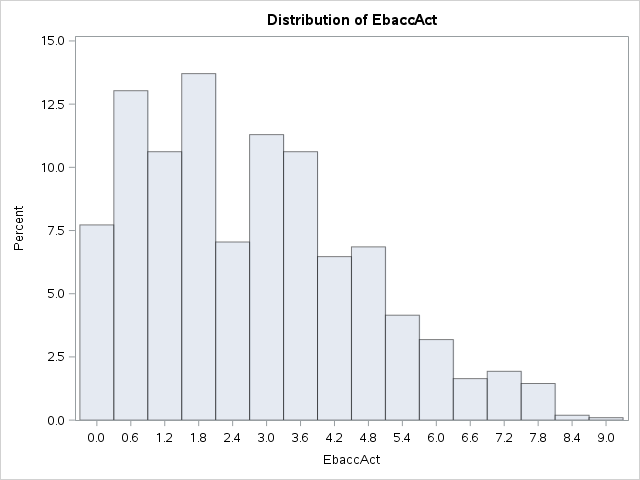
1. **MathsDiff**



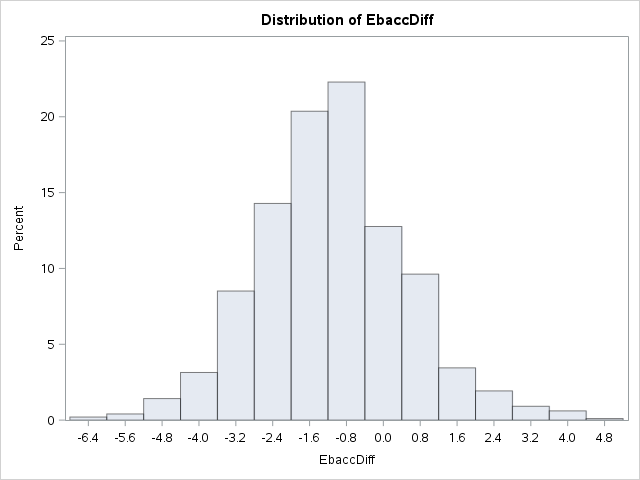
1. **EbaccEst**



1. **EbaccAct**



1. **EbaccDiff**



# **Data Quality Report:**

The report documents the predictions from a dataset containing samples from 4 years ranging over a 1000 student grades. It contains the results in different subjects such as Math and English following each an estimate and an actual grade. The data also shows English Baccalaureate scores. There was no missing data. Most of the data was clear and provided accuracy in attaining our results.

1. **Data Preparation Report:**

* Feature Selection: Data Inclusion/Exclusion: For predicting the student’s final grade, we included the following columns: PP, EAL, SEN, HML, RE, WR, MA, Att8Est, Att8Act, EngEst, EbaccEst, EbaccAct. It was not clear how the estimates were made for the data provided but we decided that that was out of our scope for this project. We did not include the primary key, the difference between the estimated grades and actual grades because we didn’t have the final grades yet so having the difference didn’t make sense. We also left out the actual grades for English and Math. We left out Ebacc difference since it didn’t support the prediction in a good way.
* Data Cleaning/Missing Data: For data cleaning, we used the replace using MICE cleaning mode. This mode supplied a good prediction value as well as letting us leave some string values in our dataset that were important in determining student grades.
* Derived Attributes/Feature engineering (Transform &Simplify): We didn’t find a way to incorporate feature engineering into our models. Given more details and time, we may have been able to create ratios of current grades to help improve the overall prediction.
* Data Integration: We didn’t integrate any outside data. We did have to research Ebacc scores, what they meant, and how they were derived. We learned that the Ebacc scores were the sum of the scores from students, divided by the number of students taking the exam. We thought that this was important to include because the level of education may vary on the or even the professor. Using previous data from similar groups may help determine/predict future grades. We thought about using

1. **Data Modeling Report:**

* Regressions/Classifications
* For Att8Act and EngAct we used the Bayesian Linear Regression Model. The regularization weights were kept at 1 for both variables. For MathsAct, we used a Linear Regression Model. The L2 regularization weight was kept at 0.001 and the random number seed was 12345.

1. **Results and Recommendations:**

**Below are the R squared and RMSE scores for Att8Act, MathsAct, and EngAct**

|  |  |  |
| --- | --- | --- |
| Att8Act | | |
| Random Seed: 12345 | USING MICE |  |
| **Type of Regression** | **R Squared** | **RMSE** |
| Boosted Decision Tree | 0.83601 | 0.655149 |
| Bayesian Linear Regression | 0.87018 | 0.582911 |
| Decision Forest Regression | 0.851354 | 0.629174 |
| Linear Regression | 0.868711 | 0.586198 |
| Neural Network Regression | 0.849598 | 0.62742 |
| Poisson Regression | 0.803785 | 0.716635 |

|  |  |  |
| --- | --- | --- |
| MathsAct | | |
| Random Seed: 12345 | USING MICE |  |
| **Type of Regression** | **R Squared** | **RMSE** |
| Boosted Decision Tree | 0.630482 | 1.159188 |
| Bayesian Linear Regression | 0.718707 | 1.011382 |
| Decision Forest Regression | 0.649859 | 1.127998 |
| Linear Regression | 0.722665 | 1.004242 |
| Neural Network Regression | 0.718602 | 1.01157 |
| Poisson Regression | 0.650253 | 1.127749 |

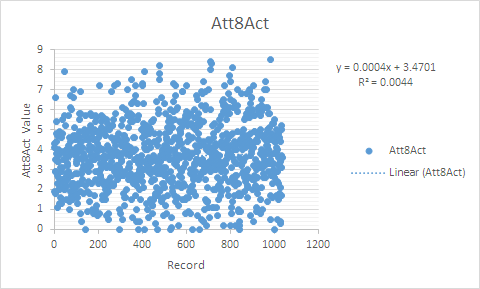
|  |  |  |
| --- | --- | --- |
| EngAct | | |
| Random Seed: 12345 | USING MICE |  |
| **Type of Regression** | **R Squared** | **RMSE** |
| Boosted Decision Tree | 0.457972 | 1.325066 |
| Bayesian Linear Regression | 0.548831 | 1.208917 |
| Decision Forest Regression | 0.527045 | 1.255107 |
| Linear Regression | 0.545139 | 1.213853 |
| Neural Network Regression | 0.544262 | 0.544262 |
| Poisson Regression | 0.511437 | 1.258019 |

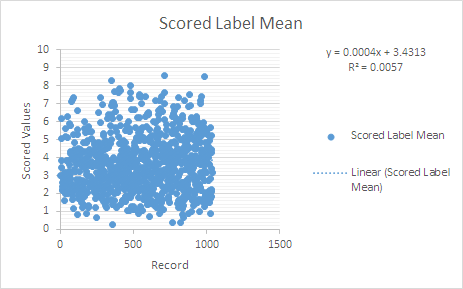
Below are charts with scatter plots of all the records.

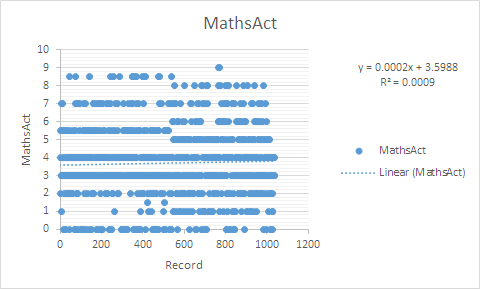
Our findings show that our Azure models had the Att8Act as the best predicted variable. According to the scatter plots below, EngAct has the most resemblance. Looking at the predicted values in the Excel file, there are some predicted values that are spot on and others that are completely off. For further research, we would investigate any similarities between data that has bad prediction values. From here, we could decide to what we would do with the data, like binning for example or clipping depending on the cause of bad prediction.

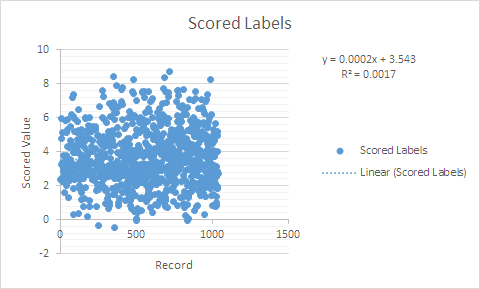
We started off this project thinking that we should do separate machine learning models. Then we moved towards thinking that we should predict Att8Act because that had the highest r squared value, then use that value to predict MathsAct followed by EngAct. However, the R squared and RMSE values would be very bad since we’re already using predicted data. We scrapped this and went back to our original idea which is shown in this paper.

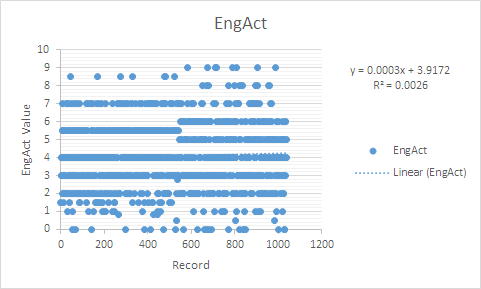
So, can this predictive model be used to predict the Att8Act? Not entirely. We have some predictions that are spot on, but we also have a difference that would be substantial enough in a real-world application where it shouldn’t be depended on. There is potential for improvement with having more complete data and more time to figure out the details of what causes the poor predictions to be that wrong. I would recommend more research and clarification on the data provided, then approaching it with the same model as we did and some additional models depending on how the data is prepared.

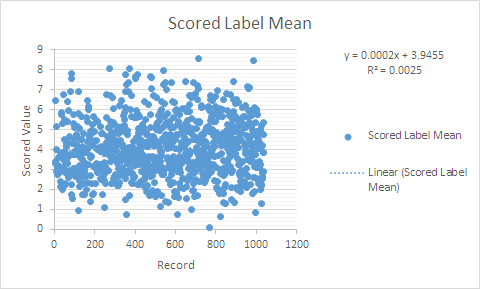












1. **References:**

* <https://www.kaggle.com/daviddraper1518/predicting-student-grades#AllResults.csv>
* <https://app.myeducator.com/reader/web/1561b/>