**COMPONENT 3 - REPORT**

All the data comes from the UK Government, most available from the website below:

<https://www.gov.uk/>

**What are the factors that decide the OFSTED(Office for Standards in Education, Children's Services and Skills)Ratings of schools???**

Why is OFSTED rating important !!

According to a study, Ofsted ratings are the third most important factor to parents when choosing a school, after location and suitability to the child’s needs.(Source Google)

If a school can improve its Ofsted rating it could affect the number pupil joining the school. We are going to find out features / factors which can help schools improve their OFSTED ratings.

# Ofsted rating

Ofsted opens in a new windowrates schools and colleges on a four point scale.

|  |  |
| --- | --- |
| 1 | Outstanding |
| 2 | Good |
| 3 | Requires Improvement |
| 4 | Inadequate |
| - | No Ofsted assessment published |

However in the data provided to us, there are values ‘Unknown’ & ‘Satisfactory’.

On checking further it was found that schools having ‘Satisfactory’ in the dataset , actually have ‘Unknown’ on the website. A possible reason could be as the data was collected in past it has been updated. This raises a concern on reliability of the data provided to us.

Talking about the data that is provided there are 5 datasets which are used for this study.

* england\_ks4-pupdest
* england\_ks4provisional
* OFSTED Ratings
* england\_abs
* england\_census
* england\_school\_information

Each school is assigned URN ID to it.

We can validate data for any school by searching the URN id of that school in the link mention below:-

<https://www.compare-school-performance.service.gov.uk/schools-by-type?step=default&table=schools&region=all-england&for=secondary>

We have multiple datasets containing different sort of information , for eg. england\_school\_information has information about School ID , Local authority , Gender (Mixed/ Single Sex school) and religious character.

We observe that all the datasets that are provided to us can be joined together on basis of URN (Unique reference / School ID)

We have left joined england\_ks4-pupdest , england\_ks4provisional, OFSTED Ratings, england\_abs, england\_census with england\_school\_information dataset bringing in different kind of information for each URN.

On joining the tables we found out that except for 4 columns every column has some missing values.

For some columns there are about 35% missing data & also in some features missing data is for school type ‘Independent schools’ , which shows there is gap in data collection moving from school to school.

As we have missing data , missing value treatment becomes an important task before moving to the machine learning task.

Approaches for missing value treatment that can be followed -

1. **Deletion**

**Deleting all the rows wherever even one value is missing for any of the column.**

**This can be done using dropna() function in pandas.**

**This approach can lead to problem where we might loose some important information.**

1. **Imputation (Mean / Mode/ Some value)**

We can impute the missing data points for each variable having missing values by looking at the group means of that variable or the values that occur most or with a constant value.

This approach can lead to overfitting when we move towards the machine learning task.

1. **Impute by Predictive Model**

Using a predictive model/ machine learning model to impute the missing values

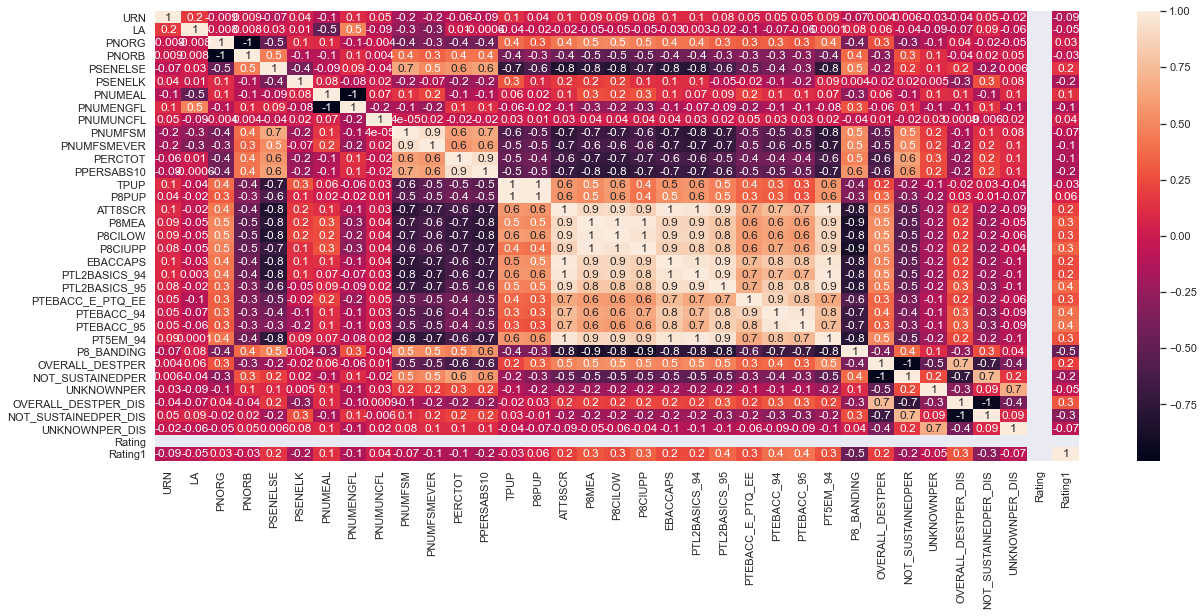
We can use decision trees / Knn to fill the missing values.

Correlation

The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0.

|  |
| --- |
| Exactly –1. A perfect downhill (negative) linear relationship |
| –0.70. A strong downhill (negative) linear relationship |
| –0.50. A moderate downhill (negative) relationship |
| –0.30. A weak downhill (negative) linear relationship |
| 0. No linear relationship |
| +0.30. A weak uphill (positive) linear relationship |
| +0.50. A moderate uphill (positive) relationship |
| +0.70. A strong uphill (positive) linear relationship |
| Exactly +1. A perfect uphill (positive) linear relationship |

We started with creating a correlation plot to find relation between different variables that are there in the dataset.



Features having positive correlation:-

* PSENELSE(% of SEN people with EHC plan)
* ATTR8SCR(Average Attainment 8 score per pupil)
* P8MEA(Progress 8 measure after adjustment)
* P8CILOW & P9CIUPP
* EBACCAPS(Average EBACC APS score)
* PTL2BASIC\_94 & PTL2BASIC\_95(% achieving English Baccalaureate with 9-4 passes & 9-5 passes)
* PTEBACC\_95

Feature having negative correlation:-

* PSENELK(% of eligible people with SEN support)
* PPERSABS10(% of enrolment who are persistent absentees)

If the schools can improve the features that are positively correlated then it might help them improve their ratings.

Also, our research points out at some features which are negatively correlated which means these features negatively affect the rating of a school. For eg. Higher numbers of persistent absentees mean it his highly likely that the school will have a poor rating.

The conclusion would be there are not many features which are strongly correlated so we cannot strongly say that increasing values for definitely increase the rating but as there are features which have moderate correlation means that improving on those features will certainly have some positive affect on the rating of the schools.

For eg. Increasing the P8 score will have a positive effect on the rating of the school