**Assignment ― Feature Engineering**

The Responses to Educational Disruption Survey (REDS) collected questionnaire data from Grade 8 students, their teachers, and school administrators to evaluate the effectiveness of several countries’ strategies to continue education during the pandemic. It was administered between December 2020 and July 2021. The surveys asked about both positive and negative effects of the education policies on students and school staff.

**Datasets**:

< Student Questionnaire Data.csv > Student data from the United Arab Emirates (UAE)

< Teacher Questionnaire Data.csv > Data from teachers of the sampled students

< School Head Questionnaire Data.csv > Data from administrator of the sampled students’ school

**Codebook**: See pages 3 to 4 (below) for the definition and measurement scale of each feature column.

**Label (Outcome) Column**: *IS18C\_inc\_plan\_capacity* in the Student data.

1. The identifier columns in these datasets are named ‘IDSTUD’ (unique student identifier) and ‘IDSCHOOL’ (school identifier). Which ID(s) are present in each dataset?

Ans. Here's the presence of identifier columns in each dataset:

Student Dataset: Contains both IDSTUD (student ID) and IDSCHOOL (school ID).

Teacher Dataset: Contains IDTEACH (teacher ID) and IDSCHOOL (school ID).

School Dataset: Contains only IDSCHOOL (school ID).

1. The Student dataset does not contain any identifier to connect individual students to their teachers. To include students and teachers in the same analysis, one solution is to **aggregate** the Teacher data within each school.

Choose two columns in the Teacher dataset that you believe might predict students’ success learning during the pandemic. Aggregate each column as a count, mean, median, or another summary statistic for teachers in each school. Which summary statistic did you use? Explain why.

Ans.

Two columns from the Teacher dataset that could predict students’ success learning during the pandemic are

1. IT06A\_teach\_qual: This column describes Extent to which Grade 7 teacher maintained their quality of teaching. This attribute maintain teaching quality during pandemic can influence how well students adapted and continue learning. I can use Mean for summary statistics.

The mean captures the general quality of teaching in a school. If most teachers felt they maintained teaching quality, students in that school were likely supported better.

2. IT05F\_indiv: This column describes *change in time spent assisting students on a one-on-one basis.* During pandemic personalized support could have been crucial during remote or disrupted schooling. We can use Median as summary statistic.

The median reduces the impact of outliers in case one teacher significantly increase/decrease support. It will better reflect the experience across teachers in each school.

1. **Merge** the Student, Teacher Aggregate, and School data using the identifier columns to make a larger dataset. Report the number of rows and columns in the merged dataset.

Ans. Please check below git location for merging dataset in python.

Merged data has 2988 rows and 36 columns.

1. Conduct an **initial analysis** of the dataset that examines missing data rates, univariate descriptive statistics, bivariate correlations, and outlier observations (rows). Summarize any noteworthy results.

Ans.

**Missing Data Rates**

Most columns have very low missing data. Below are the most 5 columns have most missing data

`IS25H` (More worried than before): 7.46%

`IS25G` (Did not feel like contacting friends): 7.33%

`IS25D` (Felt lonely): 7.33%

`IS25F` (Felt angry more often than usual): 7.30%

`IS25E` (Got upset easily): 7.26%

**Univariate Descriptive Statistics**

We have total of 2,988 student records in merged dataset

Age (`ASDAGE`) ranges from 10 to 17 years, with a mean age of 13.4

Gender (`SEX\_female`) is fairly balanced of 51.6% female.

Some important indicators below:

`IS14A` (hard to understand schoolwork) has a mean 2.37

`IS21\_teach\_sup` (teacher support) averages 2.87 out of 4,

`TeachQual\_Mean` is quite high at 3.15

**Bivariate Correlations (Top 5 Columns)**

Correlations among the first few variables (like `IDSTUD`, `IDSCHOOL`, `TOTWGTS`) are trivial and mostly structural.

`IS18C\_plan\_capacity` and ASDAGE shows low correlation with other numeric features — implying its predictors might be complex or non-linear.

**Outliers**

Using IQR (interquartile range) method, 1,497 rows (50%) had at least one outlier

**Noteworthy Findings**

* Mental health and well-being indicators (`IS25` series) have the highest missing rates and may need focused handling.
* The dataset shows good coverage and balance overall.

1. In your opinion, is it useful to **impute** missing values in this dataset? Is it useful to **augment** the dataset by adding new rows? Explain why.

Ans.

Yes, in most cases, imputing missing values is useful. When the percentage of missing values is low like

8% in most columns. Also, when we want to retain as much data as possible for modeling and analysis.

Many machine learning models can't handle missing data directly. Imputation avoids losing valuable rows that are mostly complete.

It’s not advisable to augment the dataset by adding new rows. The data set we are dealing with is survey response. So artificially generating new students would risk biasing the distribution.

1. Conduct **dimensionality reduction** of the ‘extent of difficulty experienced’ item set.
   1. Is it necessary to transform any data columns before this analysis? If yes, indicate how the data should be transformed, and justify your decision. If no, explain why.
   2. Summarize any noteworthy results.

Ans.

Yes, transformation is recommended before dimensionality reduction.

We can conduct dimensionality reduction on the “extent of difficulty experienced” item set in your dataset.

Below are the columns from student data set where we can apply dimensionality reduction

IS14A, IS14C, IS14K

IS17A, IS17C, IS17E

IS25A to IS25J

For example, IS25D (felt lonely), IS25E (got upset), IS25F (angry) columns could be labeled "Emotional stress".

Another might work like IS14A (hard to understand), IS14K (concentration), IS17A (quiet space)

could be labeled "Learning environment"

If many variables are strongly correlated and captured in a single component,

we can reduce noise and improve model interpretability by using only those components.

Here we can't see those correlated data set.

For transformation we can apply standardization (z score normalization).

1. Conduct **feature selection** using a decision-tree model.
2. Is it necessary to transform any data columns before this analysis? If yes, indicate how the data should be transformed, and justify your decision. If no, explain why.
3. Summarize any noteworthy results.

Ans.

No major transformations are strictly necessary because decision tree can handle categorical and ordinal variables. It can also can handle missing values. We can Drop or impute missing values in the predictor or target columns. We can use a decision tree classifier to identify which features are most important in predicting the target.

Please see the below git url for applying decision tree classifier.

Here are the top 5 most important features identified by the decision tree model in predicting the

student's capacity to plan schoolwork (IS18C\_plan\_capacity):

TOTWGTS - Student's sampling weight

ASDAGE - Student's age

TeachQual\_Mean - Avg. teacher-reported ability to maintain teaching quality

IndivAssist\_Mean - Avg. teacher-reported time spent on 1-on-1 student help

IS21\_teach\_sup - How supported the student felt by teachers

Summary

* Student's age and parent education level are strong demographic predictors of planning ability.
* Teacher support and quality (both directly reported and through student perception) are critical.

1. **Predict** the odds that individual students reported increased capacity to plan their schoolwork (i.e., treat ‘IS18C\_plan\_capacity’ as the label column) using multinomial logistic regression.
2. Give a brief rationale for the feature columns that you chose to include in the prediction model.
3. Is it necessary to transform any data columns before this analysis? If yes, indicate how the data should be transformed, and justify your decision. If no, explain why.
4. Summarize any noteworthy results.

Ans.

I selected features that reflect a student’s demographics, learning environment, emotional state, and support systems during the pandemic:

|  |  |
| --- | --- |
| **Feature** | **Reason for Inclusion** |
| ASDAGE | Older students may have more maturity or responsibility in planning. |
| SEX\_female | To capture potential gender differences in learning responses |
| prnt\_ed\_M | Parental education often correlates with learning support at home. |
| IS14A | Difficulty understanding schoolwork impacts ability to plan. |
| IS14K | Trouble concentrating makes planning harder. |
| IS17A, IS17C | External factors like quiet space or caring for siblings affect planning. |
| IS21\_teach\_sup | Feeling supported by teachers can boost planning confidence. |
| IS09\_comm\_modes | More communication channels may aid in organizing schoolwork. |
| TeachQual\_Mean, IndivAssist\_Mean | Quality of teaching and 1-on-1 support can influence planning skills. |

Yes, it is necessary to transform the data before logistic regression. Most of our variables are ordinal or continuous, but on different scales like ‘ASDAGE’ in years, ‘IS14K’ on a 1–4 Likert scale, ‘TeachQual\_Mean’ as an aggregated score. We can apply Z-score standardization to scale features using sklearn library from python.

Using logistic regression with carefully selected and standardized features helps identify the **key factors** that influence students' ability to plan during disruptions like COVID-19.

|  |  |  |
| --- | --- | --- |
| **Sampling Design Variables** | | **Codebook** |
| IDSTUD | Student ID | Unique row identifier; nominal |
| IDSCHOOL | School ID | Nominal (unordered) categories |
| TOTWGTS | Student sampling probability weight variable | Approx. continuous |
| **Student Variables** | |  |
| IS18C\_plan\_capacity | Outcome: My capacity to plan the completion of my schoolwork. | 1: Decreased  2: Did not change  3: Increased |
| ASDAGE | Age | Years |
| SEX\_female | Indicator: Student is female | 0: No  1: Yes |
| IS34\_lang | Specific language spoken at home | Nominal (unordered) categories  1: Arabic  2: English  3: Persian  4: Urdu |
| prnt\_ed\_M | Mean education level of student’s parents/guardians derived from IS1G38 and IS1G40 | 1: Did not complete [ISCED level 2]  2: [ISCED level 2]  3: [ISCED level 3]  4: [ISCED level 4 or 5]  5: [ISCED level 6, 7 or 8] |
| IS17G\_miss\_meals | Frequency missed meals due to loss of school food program | 1: Never or hardly ever  2: Sometimes  3: Most of the time  4: Always |
| IS01\_prop\_sch\_time | Proportion of time spent in school during disruption | 1: None, attended all lessons remotely  5: Continued to attend all lessons at school |
| IS09\_comm\_modes | Count of distinct student-teacher communication modes used at least ‘sometimes’ | 0 to 4 |
| IS21\_teach\_sup | Extent student felt their teachers were supportive. Mean of four rating scale items. | 1 to 4 |
| IS18E\_adult\_help | Change in frequency of receiving schoolwork help from an adult at home during disruption | 1: Decreased during the [COVID-19 disruption]  2: Did not change during the [COVID-19 disruption]  3: Increased during the [COVID-19 disruption] |
| **Feature Set**:  Extent of difficulty experienced by student while learning at home | IS14A [hard to understand schoolwork]  IS14C [able to complete schoolwork without help]  IS14K [difficulty concentrating]  IS17A [quiet space to work]  IS17C [had to care for sibling(s)]  IS17E [adequate materials]  IS25A [exercised more than before]  IS25B [did more hobby activities, e.g., scouting, sports]  IS25C [felt fit and healthy]  IS25D [felt lonely]  IS25E [got upset easily]  IS25F [felt angry more often than usual]  IS25G [did not feel like contacting friends]  IS25H [more worried than before]  IS25J [did not sleep as well] | Sets 14 and 25:  1: Strongly disagree  2: Disagree  3: Agree  4: Strongly agree  Set 17:  1: Never or hardly ever  2: Sometimes  3: Most of the time  4: Always |
| **Teacher Variables** | |  |
| IT25\_yrs\_exp | Grade 7 teacher years of teaching experience | 1: Less than 1 year  2: 1–2 years  3: 3–5 years  4: 6–10 years  5: 11–20 years  6: Over 20 years |
| IT06A\_teach\_qual | Extent to which Grade 7 teacher maintained their quality of teaching | 1: Strongly disagree  2: Disagree  3: Agree  4: Strongly agree |
| IT05F\_indiv | Change in time spent assisting students on a one on-  one basis | 1: Substantially decreased  2: Decreased to some degree  3: Did not change  4: Increased to some degree  5: Substantially increased |
| IT06D\_indep | Extent to which the materials provided to students enabled them to work independently | 1: Strongly disagree  2: Disagree  3: Agree  4: Strongly agree |
| **School Variables** | |  |
| IP34\_locale\_size | Size of city where school is located | 1: Fewer than 3,000 people  2: At least 3,000 but fewer than 15,000 people  3: At least 15,000 but fewer than 100,000 people  4: At least 100,000 but fewer than 1,000,000 people  5: 1,000,000 or more people |
| IP1G34A\_public | School is public / government owned | 0: No  1: Yes |
| IP03\_tech | Count of technology (computer, Internet) provided to each student at home | 0 to 2 |
| IP1G35C\_prop\_disadvan | Proportion of socio-economically disadvantaged students | 1: Less than 5%  2: 5-10%  3: 11-25%  4: 26-50%  5: More than 50% |
| IP13D\_prnt\_sup | Extent to which school was supported by students’ parents/caregivers | 1: Not at all  2: Somewhat  3: Very well |
| IP21\_remote\_prep | Extent to which school is prepared to deliver remote instruction in future | 1: Not at all prepared  2: Not very well prepared  3: Well prepared  4: Very well prepared |