Capstone project 1 Milestone Report -Historical Enterprise Incentive scheme Business data

Client:

Australian Department of Employment

Problem to solve:

Target of our problem is to find out whether an individual planning to start a business under NEIS scheme will be successful or not - This will be useful to vet out individual business application prior to approving them and providing mentor support as part of NEIS scheme.

Dataset:

https://data.gov.au/dataset/neis

Data Wrangling:

Listed below are the observations based on data (retrieved from Excel):

1. There are around 53K+ records in excel. Only certain columns have at least 10K+ valid data inside these columns. Rest of the data involves NANs, which would require data imputation
2. There was one non-categorical column 'SV_HOURS_WORK' and this holds 8K+ records. Therefore, we have computed the mean and replaced NANs with the mean.
3. Some of the categorical features (such as SV_END_TRAIN, SV_END_MENTOR, SV_END_PROFIT etc.) have very less non-NAN records and it is difficult to classify/impute the data in these columns as each record may hold any one of categorical value. For instance, consider the columns SV_END_TRAIN Has 4 options - .,1 - No,2 - yes a little,3 -yes a lot. These options indicate whether the business ended due to poor quality training.
4. Data is then plotted to analyse the impact of various features on the 'TARGET' (column 'Success indicator' in the data set). Based on the plots/results derived:

    —> Aged groups 25 to 45 are relatively more successful than their counterparts.

    —> NSW, QLD and VIC have large number of participants in the program and NSW has a high success rate

    —> Business owners that don't belong to any particular community/ disability type appear to be the most successful ones.

    —> SV_SAT_OVERALL is the categorical feature that influences the target based on heat map analysis.

    —> Other categorical features that influence the target are: PERSONALITY_TYPE, AGE_GROUP, STATE, INDUSTRY_TYPE etc.

Inferential statistics:

As a part of inferential statistics, following have been analysed:

1. Pivot table and plots were generated to find how successful businesses are based on (i) Industry type and (ii) personality type of the participant
2. ECDF plot was created for non-categorical column 'SV_WORK_HOURS' (number of work hours per week) and as most of the records had NAN's in this column, we had to impute NANs with mean of the available values. As a result, 80% of the businesses had work hours less than or equal to 35 in the ECDF plot
3. We generated a theoretical ECDF by normalising the randomly generated values (based on the mean and Standard deviation of actual values). Theoretical ECDF showed that around 97% of the businesses can have > 50 hours of work per week.
4. A Null hypothesis test was conducted to find whether the "Number of business work hours" can be

greater than or equal to 50. A p-value of 0.4929 prompted to accept the hypothesis.

Next Steps:

Moving forward, A basic machine learning model on logistic regression will be created and analysis would be conducted with respect to the accuracy of the model and its performance.
After completion of ML Model, a final Report and presentation deck will be submited to conclude the Capstone project 1.

All the code Submissions are present under the GITHUB:
https://github.com/vsrajesh1/Capstone_Project_1