Springboard DSC

Capstone project 1

Rajesh V Subramanyam

March 2019

**Introduction**

NEIS(New Enterprise Incentive scheme) is a a commonwealth government program to help eligible Australians(Job-seekers) into self-employment. The programme includes training in small business management, a business plan, one year of income support and mentor support. It is a programme delivered by a network of 21 providers who provide individualised help for job seekers to become self employed business owners. There are different types of assistance available under this program such as 'Training', 'Rental Allowance', 'New Business Assistance', 'Business mentoring and Support' etc.

This capstone project deals with the historical business level data of NEIS between 1 July 2007 and 31 May 2017, which involves business level administrative data and Post-Program Monitoring business level survey data. Goal of this project is to determine the success of a prospective applicant based on the historical data

**Background**

NEIS was officially launched as a pilot scheme by the then Minister for Employment and Industrial Relations, the Hon Ralph Willis AO. At the 1985 launch, the then Minister said that the joint New Business Enterprise Scheme (now known as NEIS) was innovative in its provision of both start-up capital and income support, and that it will ‘provide unemployed people with comprehensive support for their entrepreneurial schemes which would otherwise not be available.’ One factor in the Scheme’s continued success has been the expertise and experience that contracted NEIS providers have provided over a sustained period. Many of the NEIS providers have also delivered other Australian, State or Territory government small business related programmes and are long term Registered Training Organisations (RTOs) for the delivery of NEIS small business training. The variety of businesses that have been established through NEIS over the years has been impressive, from nutrition education to a live music venue that has hosted a raft of international acts.

**Dataset / Features**

We have 2 excel documents as part of the dataset - 1) Mail Excel and 2) Data dictionary. While the main excel contains actual data required for our analysis, the 'Data Dictionary' contains the metadata of the features, description and the possible values for the categorical features. Based on the data points present in the Main Excel i.e. Some of the columns are missing values (NAN), which makes these columns difficult to be considered for analysing the target. There are 53,000+ records in total and Columns having atleast 10K+ non-NAN records are listed below. (1) business\_id (2) start\_date (3) end\_date (4) exit\_reason (5) successful (6) industry\_type (7) state (8) metro (9) age\_group (10) gender\_cd Apart from the above columns, we have some indicators that denote the personality type/ community of a person. Rather than considering multiple indicator columns, we will add one column "Personality\_type" to the dataframe and that will explain the personality/community type of a business owner (11) Personality\_type

**Features that need imputation to handle the NAN records**

The non-categorical column 'SV\_HOURS\_WORK' seems to have 8K records. This column was imputed by replacing NANs with the mean of the values present in this column.

**Analysis of the dataset**

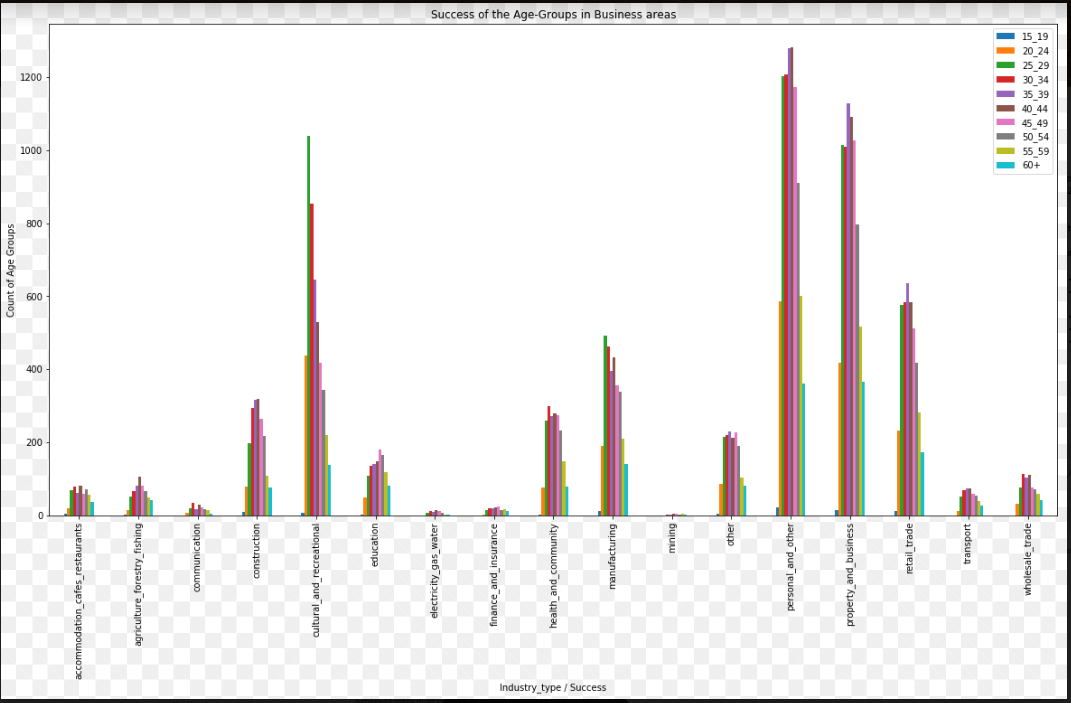
***Influence of Age group on the success of the business***

Bar chart was plotted between the feature 'Age\_group' and the target "Success\_indicator" to find the influence of Age on the success of the Business. Listed below are the observations:

1) Age groups between 25 to 45 are relatively more successful than those of other age groups

2) The above age groups are very successful in the industry types 'Property\_and\_business' and 'personal\_and\_other'

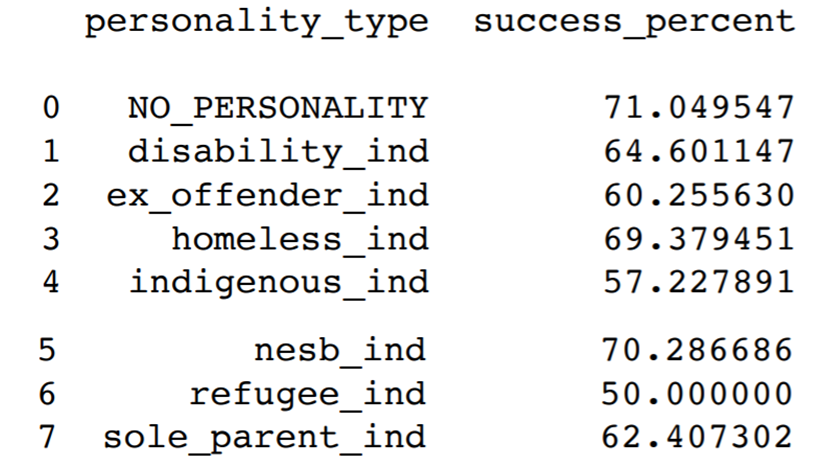
3) All the age groups were least or not so successful in industry types 'mining'

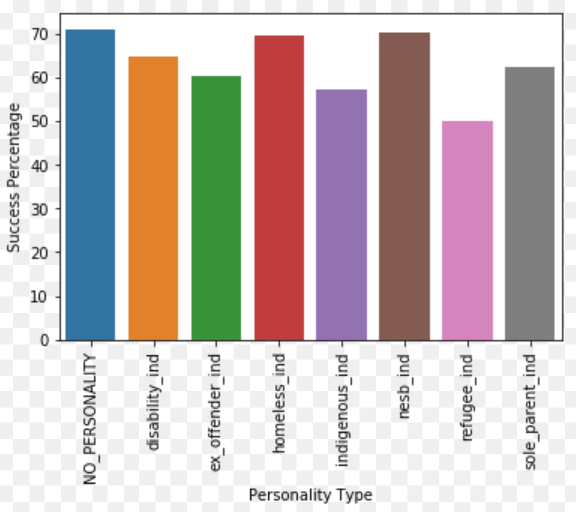


***Influence of Community/Personality Type on the success of the business***

1) Businesses started by Refugees are successful only 50% of the time- Department might have to allocate additional funds / training before allowing the refugee participants to start their own business.

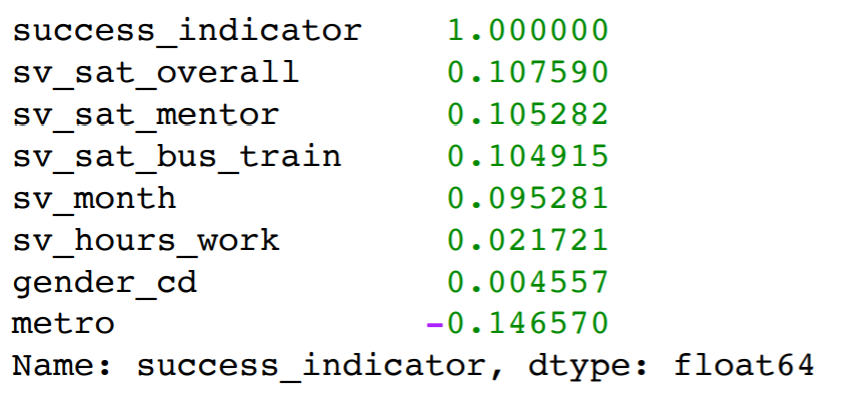
2) Although the Australian government has introduced policies/programs to encourage Aboriginal/indigenous participants, they are 57% successful, which shows that government will have to undertake additional measures.

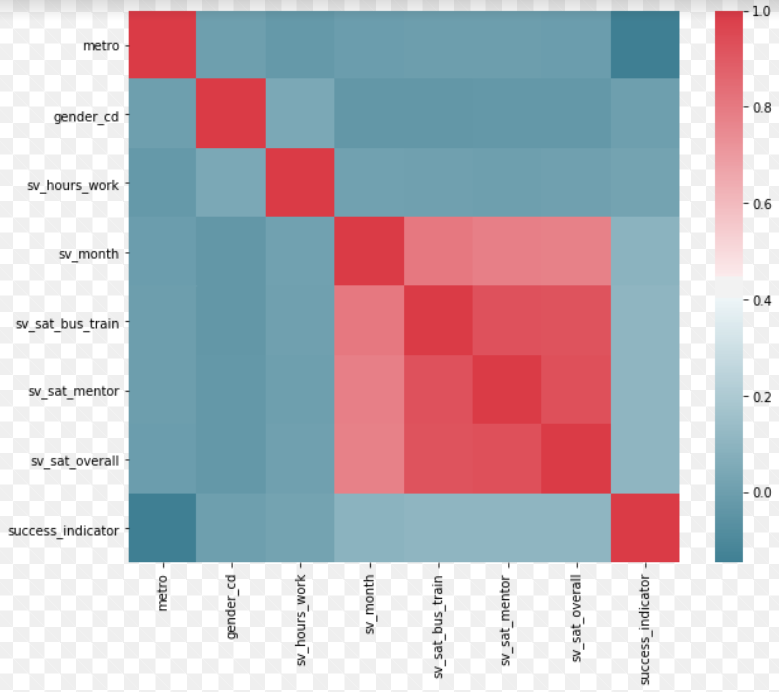




***Heatmaps of the features and their correlation to the Target:***

As per the Correlation results, only certain features are positively correlated such as sv\_sat\_overall, sv\_sat\_mentor, sv\_sat\_bus\_train and sv\_hours\_work. All other features are either Close to ZERO/ negatively correlated. But none of the features appear to be very strongly correlated with the target "SUCCESS INDICATOR"

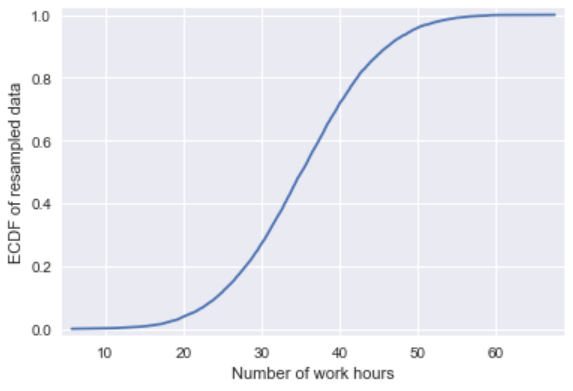




**Inferential Statistics**

***ECDF of non-categorical feature SV\_WORK\_HOURS (Number of employee work hours per business)***

Non-categorical variable 'SV\_WORK\_HOURS' is one of our features, which has a positive correlation with our Target variable 'SUCESS\_INDICATOR'. On conducting the ECDF for this feature, we found that 80% of the businesses have work hours less than or equal to 35 hours. But this was due to imputation performed on this feature to remove the NAN values. Therefore, we perform bootstrapping on this feature with random sampling and plot the theoretical ECDF (shown below). According to the theoretical ECDF distribution, 97% of the businesses have > 50 hours of work hours per week. This is very different from the actual result (of 35 hours per week)



***Null Hypothesis: Mean working hours of business is greater than or equal to 50?***

A null hypothesis was performed on the feature SV\_WORK\_HOURS to check whether the hypothesis can be accepted. A P value of 0.49 is obtained and is greater than the significant level(0.05,0.01,0.001). This suggests that we accept (the NULL hypothesis) that the mean working hours of businesses can be greater than or equal to 50

***NULL Hypothesis: Check whether there is acceptable level of participation by Aboriginal people in the program***

As part of this hypothesis test, we wanted to check whether the aboriginal participants are participating in the same ratio as that of other personalities. A P-Value of 0.48 implies that the hypothesis can be accepted. It is likely that there is equal participation from all communities/ personality types involved in the program.

**Machine Learning: Building a Logistic Regression Model**

We are building a regression model to predict the success of a prospective applicant. For this purpose, the historical data is split into two halves. One for training the model and the other half to be used for testing the same.

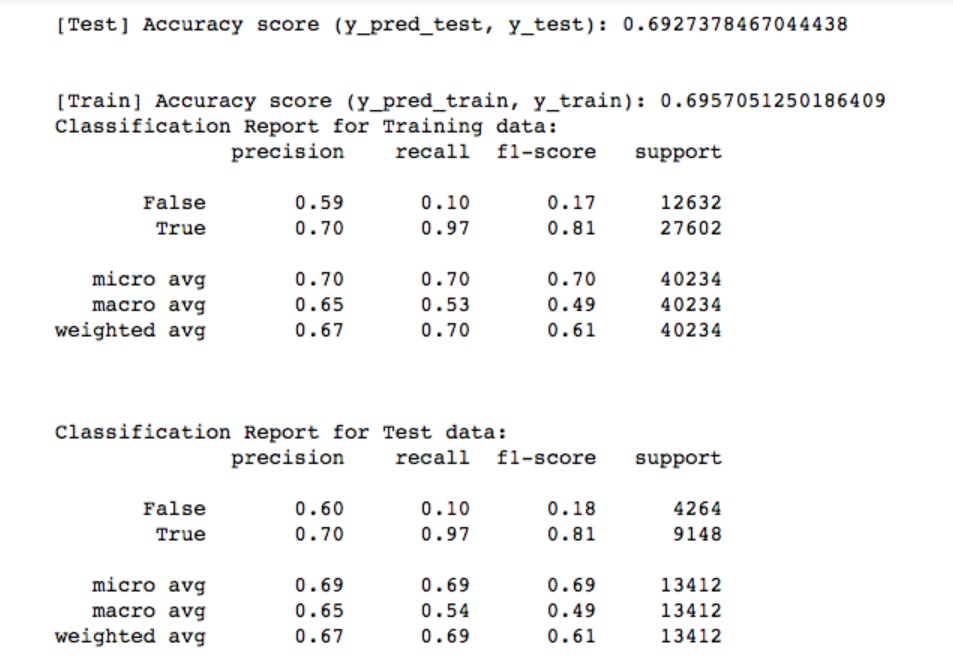
Listed below are the features used to construct this model:

1. 'industry\_type' 2) 'state' 3) 'gender\_cd' 4) 'sv\_hours\_work' 5) 'metro' 6) 'age\_group'

7) 'indigenous\_ind' 8) 'ex\_offender\_ind' 9) 'nesb\_ind' 10) 'refugee\_ind'

11) 'disability\_ind' 12) 'homeless\_ind' 13) 'sole\_parent\_ind'

After conducting the logistic regression on the data, we received an Accuracy score of 0.69 (for training/ test datasets) Classification results are listed below:



***Based on the above results:***

1) Both training and test accuracy scores are almost the same and so there is no OVERFITTING

2) There is no large gap between any of the observed metrics - PRECISION, RECALL & F1-SCORE.

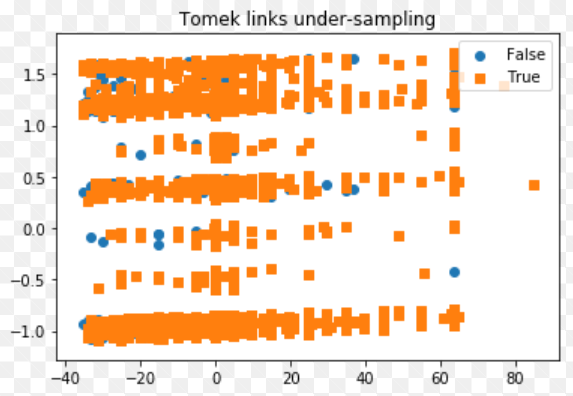
3) Dataset comprises of more successful businesses as against the failed ones, which is evident from the classification report i.e. More 'TRUE' than FALSE.

***Imbalanced classes:***

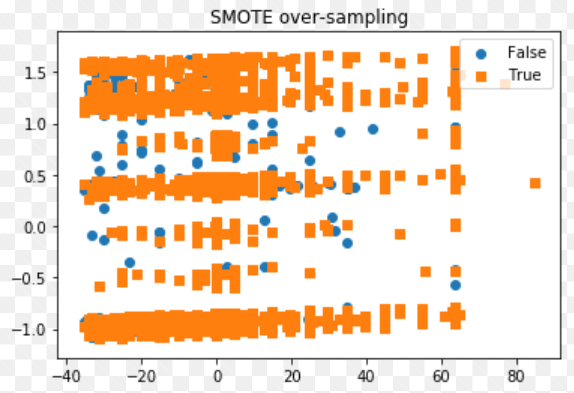
Based on the results in the previous section, we could deduce that the dataset involves i.e. dataset involves 80% of successful businesses and less than 20% of failed ones.

In order to address this issue, Over-sampling and under-sampling techniques present in the python library 'imblearn'. Plots (shown below) are created using different methods available under IMBLEARN. Listed below are some of these plots and their analysis:

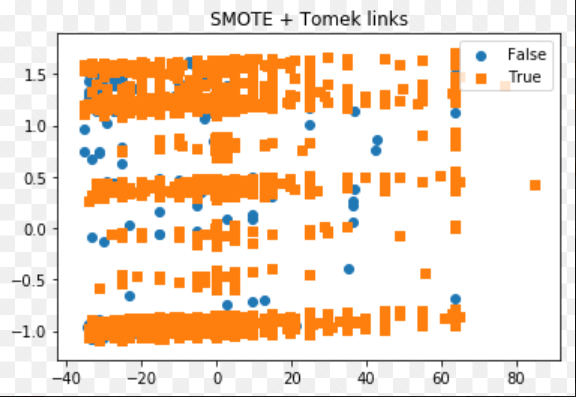
1. Under sampling using TomekLinks



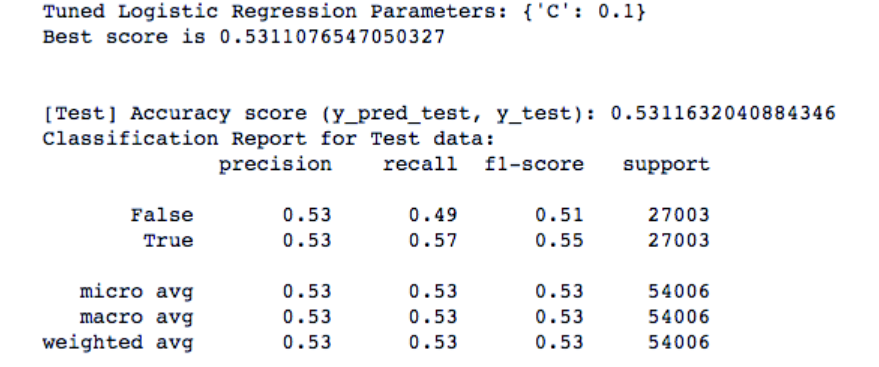
1. Over sampling using SMOTE



1. Over sampling using SMOTE + TOMEK



***Accuracy score and classification report is generated for data fitted using SMOTE***

******

Based on the above results:

1) We find that the classification scores have greatly improved. There is equal percentage of data points involving both TRUE, FALSE

2) But, the accuracy score had reduced to 0.53 from 0.69. In order to ensure that scores along with accuracy, we may have to collect more data so as to ensure that the occurrences of 2 classes (TRUE/FALSE) are balanced.

**Conclusion & Recommendation to the client:**

In order to determine the target ‘SUCCESS\_INDICATOR’, we performed logistic regression model involving training and test data set. However, one class of the target (TRUE) is more prevalent in the dataset than the other class (FALSE) – This was observed based on the classification results. In order to improve the involvement of both the classes, we conducted over-sampling/under-sampling techniques present under the library IMBLEARN. Although there is a significant improvement in recall, f-1 score (classification results), the accuracy of model prediction (of Target) has reduced from 0.69 to 0.53. A summary of classification results observed for the ML Techniques used in this project are listed below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML Technique | Accuracy score | Class of the Target | Precision | Recall | F1-Score |
| Logistic Regression | 0.69 | FALSE | 0.60 | 0.10 | 0.18 |
| TRUE | 0.70 | 0.97 | 0.18 |
| SMOTE + TOMEK | 0.53 | FALSE | 0.53 | 0.49 | 0.51 |
| TRUE | 0.53 | 0.57 | 0.55 |

Future course of action would involve collecting more data from NEIS. Still we could have more occurrences of one class (TRUE) of target as the NEIS has strict admission criteria for participants. However, this would at least give us more meaningful statistics/ results.

Listed below are the recommendations for the client based on the results:

1. Machine learning model achieved a prediction accuracy of 70%. Therefore, this ML model requires further tuning and until then client may only use it as an additional reference rather than taking decisions based on this result.
2. We had several data points and features. However, most of the features had NANs and required imputation. If we can avoid having these NANs in the first place by capturing the necessary values, then accuracy/ prediction of ML model can be improved.
3. In order to address the problem of imbalanced classes, ML techniques such as SMOTE, TOMEK etc were used in previous section. However, it is observed that the accuracy score had reduced to 0.53 from 0.69 although the classification score has greatly improved. In order to ensure good classification scores along with accuracy, we may have to collect more data so as to ensure that the occurrences of 2 classes (TRUE/FALSE) are balanced.