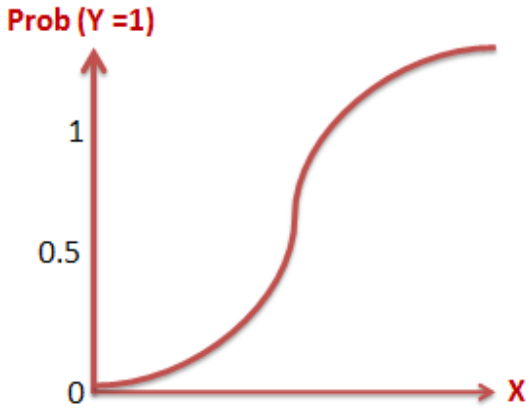

Apply Logistic Regression to Analyze Singapore Workplace Injury Data

EBS5101 Foundation of Business Analytics – Assignment 1

Submitted by:

Ma Min(A0163305N), Muni Ranjan(A),

Pradeep Kumar (A0163453H), Zheng Weiyu (A0163412R)

CONTENTS

Objective

The objective of this report is to explain the team work done to apply data exploration learning technique. We have selected the data "Workplace Injury by types" provided by Singapore government. We would like to identify the relationship between different factors provided in the data. We want to find out if there is an independent variable which could be predicted based on one or more dependent variable.

Below is the quick snapshot of data:

	year	degree_of_injury	industry	sub_industry	incident_type	incident_agent	incident_agent_sub_type	no._of_injuries
2	2011	Fatal	Community, Social & Personal Services	Repair & Maint	Caught in/ betw C	Vehicles	Vehicles - Motor vehicles	1
3	2011	Fatal	Community, Social & Personal Services	Repair & Maint	Falls - Slips, Trips	Vehicles	Vehicles - Motor vehicles	1
4	2011	Fatal	Construction	Civil Engineerir	Collapse/Failure c	Others	Others - Furniture and Fitt	1
5	2011	Fatal	Construction	Civil Engineerir	Struck by Moving	Lifting Equipment In	Lifting Equipment Includin	1
6	2011	Fatal	Construction	Civil Engineerir	Struck by Moving	Pressurised Equipm	Pressurised Equipments -	1
7	2011	Fatal	Construction	Construction of	Caught in/ betw C	Lifting Equipment In	Lifting Equipment Includin	1
8	2011	Fatal	Construction	Construction of	Caught in/ betw C	Vehicles	Vehicles - Excavators	1
9	2011	Fatal	Construction	Construction of	Cave-in of excava	Others	Others	1
10	2011	Fatal	Construction	Construction of	Collapse of formw	Physical Workplace	Physical Workplace - Form	1
11	2011	Fatal	Construction	Construction of	Crane-related	Lifting Equipment In	Lifting Equipment Includin	2
12	2011	Fatal	Construction	Construction of	Electrocution	Others	Others - Electrical Installat	1
13	2011	Fatal	Construction	Construction of	Falls - Falls from H	Means of Access	Means of Access - Ladders	1
14	2011	Fatal	Construction	Construction of	Falls - Falls from H	Means of Access	Means of Access - Others	1
15	2011	Fatal	Construction	Construction of	Falls - Falls from H	Physical Workplace	Physical Workplace - Struc	2
16	2011	Fatal	Construction	Construction of	Falls - Slips, Trips	Physical Workplace	Physical Workplace - Form	1
17	2011	Fatal	Construction	Construction of	Struck by falling o	Others	Others - Ceramic Items	1
18	2011	Fatal	Construction	Specialised Cor	Collapse/Failure c	Others	Others - Ceramic Items	2
19	2011	Fatal	Construction	Specialised Cor	Falls - Falls from H	Physical Workplace	Physical Workplace - Roof	2
20	2011	Fatal	Construction	Specialised Cor	Falls - Falls from H	Physical Workplace	Physical Workplace - Struc	1
21	2011	Fatal	Construction	Specialised Cor	Struck by falling o	Lifting Equipment In	Lifting Equipment Includin	1
22	2011	Fatal	Information & Communications	Telecommunic	Falls - Falls from H	Means of Access	Means of Access - Ladders	1
23	2011	Fatal	Manufacturing	Manufacture of	Falls - Falls from H	Means of Access	Means of Access - Ladders	1
24	2011	Fatal	Manufacturing	Metalworking	Collapse/Failure c	Vehicles	Vehicles - Forklifts	1

Source: data.gov.sg

Problem Dataset

By simply loading the dataset, we get the found the following information about the data:

- There are total 8 variables provided in this dataset.
- Total number of observations are 16374
- Unique values under the **no._of_injuries** varies from **1 to 261**. This indicates that for a typical accident number of workers injured from 1 to 261
- There are 3 types of degree_of_injuries - FATAL, MAJOR, MINOR

Looking at this data we were inquisitive to know:

1. Is there any relation between single injury or group injury with other factors?
2. Can we predict based on DEGREE_OF_INJURY and other factors if two or more people were involved in the accident?

To conduct this analysis we converted the injury_count to a boolean variable

- 0: Represents 1 or 2 people involved in accident
- 1: Represents more than 2 people involved in accident

For all the attributes, an initial exploratory analysis was done. Scatter plots were used to detect unusual patterns. Since there were no null values, no reduction of data was required.

Exploratory Analysis

We first identified the major attributes which could help us create the model for predicting the group injury. For this we compared the unique values in each variable and found out the following:

- Year has no effect on our model. Hence we dropped the variable
- There are too many unique attributes in the sub types which can lead to inaccuracies of the model. Those were dropped
- Degree_of_Injury(DI), Industry(IND), and Incident_agent(IA) were found suitable

As you can see in figure (1) we plotted the different accepted variables against the injury_count. It doesn't provide a very clear picture but definitely indicates that for some factors, injury_count was quite low whereas of others had a huge count.

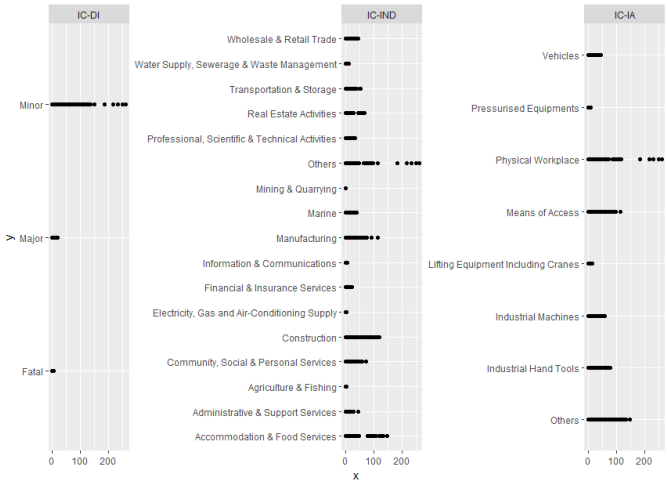

Figure 1: Scatterplot of different variables with injury_count

Determination of Key Factors

We used random sampling to select 70% of the data for training and 30% for prediction. To determine the factors influencing the injury count we used logistic regression in language 'R'. The first model summary of our data was:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3431  -0.9338  -0.6356   1.1988   2.7327

Coefficients:
                                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                                     -2.82193    0.31701  -8.902  < 2e-16 ***
degree_of_injuryMajor                            0.88313    0.30741   2.873 0.004069 **
degree_of_injuryMinor                            2.53228    0.29944   8.457  < 2e-16 ***
industryAdministrative & Support Services       -1.04507    0.11584  -9.021  < 2e-16 ***
industryAgriculture & Fishing                   -3.06381    0.72476  -4.227 2.36e-05 ***
industryCommunity, Social & Personal Services   -0.79595    0.09390  -8.476  < 2e-16 ***
industryConstruction                             0.10579    0.08882   1.191 0.233642
industryElectricity, Gas and Air-Conditioning Supply -3.49697  1.01556  -3.443 0.000574 ***
industryFinancial & Insurance Services          -1.34086    0.17472  -7.674 1.66e-14 ***
industryInformation & Communications            -2.47243    0.33651  -7.347 2.02e-13 ***
industryManufacturing                           -0.44818    0.08403  -5.334 9.62e-08 ***
industryMarine                                  -0.17320    0.11744  -1.475 0.140290
industryMining & Quarrying                     -13.21437  123.15026  -0.107 0.914549
industryOthers                                   0.24444    0.10022   2.439 0.014729 *
industryProfessional, Scientific & Technical Activities -1.08468  0.11607  -9.345  < 2e-16 ***
industryReal Estate Activities                  -0.19146    0.13388  -1.430 0.152693
industryTransportation & Storage                -0.44544    0.09684  -4.600 4.23e-06 ***
industryWater Supply, Sewerage & Waste Management -1.22084  0.16144  -7.562 3.97e-14 ***
industrywholesale & Retail Trade                -0.24829    0.10566  -2.350 0.018782 *
incident_agentIndustrial Machines               -0.22191    0.08708  -2.548 0.010824 *
incident_agentLifting Equipment Including Cranes -0.99334   0.10922  -9.095  < 2e-16 ***
incident_agentMeans of Access                    0.13496    0.09482   1.423 0.154619
incident_agentOthers                             0.13358    0.07382   1.810 0.070371 .
incident_agentPhysical Workplace                 0.42659    0.09627   4.431 9.38e-06 ***
incident_agentPressurised Equipments            -1.92927    0.19473  -9.907  < 2e-16 ***
incident_agentVehicles                          -0.03819    0.08716  -0.438 0.661271
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19800  on 16373  degrees of freedom
Residual deviance: 18185  on 16348  degrees of freedom
AIC: 18237

Number of Fisher Scoring iterations: 12
```

Figure 2: First Iteration of our model

Observations

- Fatal injury type has been filtered in 'R' output as it has low significance in predicting the group injury (greater than 2)
- We identified few more factors like industry_construction, industry_marine etc. which have high P value, hence can be dropped from the model.

After filtering out the unimportant factors we ran the iteration 2 and following are the results: