To: My Supervisor

From: PA Candidate

Date: December 13, 2018

Title: Drivers of Term Life Insurance Purchasing

Executive Summary

The main goal is to find the factors which most impact mine safety. Miners that work in dangerous conditions deserve to be paid more. This analysis uses predictive analytics in to measure the risk level after adjusting for other factors such as the type of mine, the year, and type of work performed by the miners.

The data suggests that these factors most contribute to the number of injuries sustained per 2000 worker hours.

High risk mines

- are where workers spend more than 3000 hours underground and in sand & gravel, coal, bituminous, or stone mines
- spend more than 3000 hours underground and are limestone or other mine types

The safest mines

- are where workers spend less than 3000 hours underground and more than hours 110,000 hours in the strip
- have less than 3000 hours underground, less than 110,000 hours in the strip, and less than 34,000 hours in the mill
- have less than 2957 hours underground, less than 110,000 hours in the strip, more than 34,000 hours in the mill and whose commodity is metal or non-metal

To use this in a business setting, we have created a set of rules which will predict the number of mining accidents per 2000 hours. The R^2 value of this model is 69% which means that 69% of the variation of the number of injuries can be explained by these rules. This is in the Findings section of the report. The high-level take-away from this is that dangerous mines tend to

- Be coal mines, or less risky metal or stone
- Are older (mines are becoming more risky over time)
- Have a lower seam height
- Are more crowded (have more total employee hours)

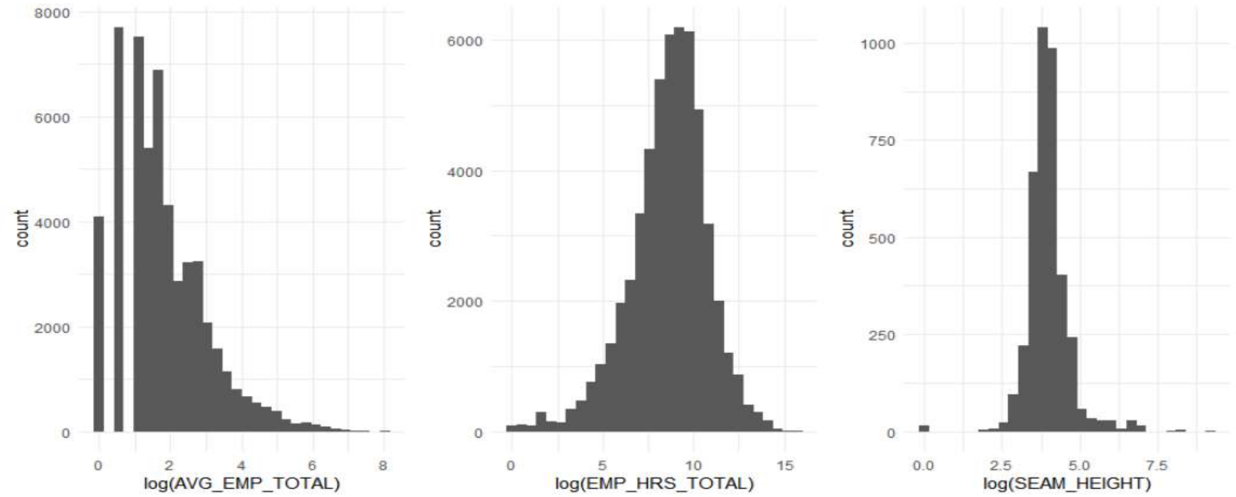
Data Exploration, Preparation, and Cleaning

The data consists of 20 variables collected from 2013 - 2016. Records from older years are likely to be less reliable and if the data collection process has changed over time then there will be bias in the data.

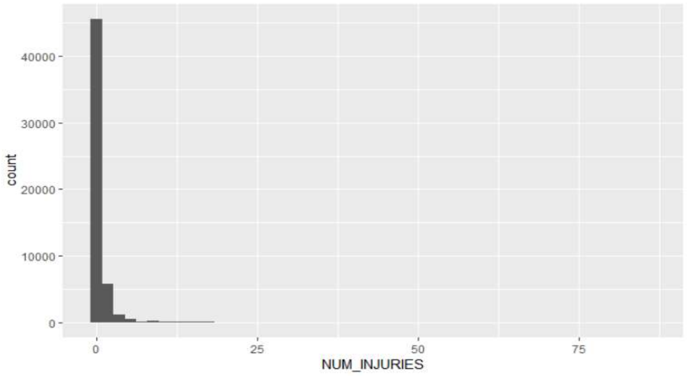
My notes from cleaning the other fields are below.

- There are about the same number of records for each year.
- There is a lot of variation in the number of records by US state.
  - PA has 3,501 mines and smaller states (or US territories) only have one.
  - The following states are being removed because they have fewer than 10 records: AS, GU, MP, NA (which is a missing value)
- There are five different types of mines. These are Coal, Metal, Nonmetal, Sand & gravel, and Stone. Most mines are sand and Gravel or Stone with a smaller percentage Coal and Metal.
- We have additional information on the primary extract of the mine. There was one record with a missing Primary field which was removed.
- The SEAM_HEIGHT field is apparently a measure of the mine dimension. There are 49,823 zero values which indicates that these values are actually missing instead of 0. I created an additional field called NA_FLAG which is 1 when the SEAM_HEIGHT is 0. This helps linear models to make sense of the non-linearity of the data.
- There are 13 mines which have a missing MINE_STATUS value. These are being removed. If I had more information I would change these to being closed. Given limited knowledge in this area I am being safe and removing these records.

The three continuous predictors AVG_EMP_TOTAL, EMP_HRS_TOTAL, and SEAM_HEIGHT were skewed and so a log transform was applied. After applying the transformation the data distributions are more symmetric which makes them easier to use in modeling. One of the assumptions of GLMs is that the covariate distributions are at least approximately normal. It is ok if these are slightly off (such as a t-distribution).



The number of injuries field is also heavily right-skewed. There are 45,600 mines (80% of the total) which have no injuries.



Feature Selection

Because we want the Union to be able to use these results in a variaty of geographic contexts, we may consider excluding the STATE field altogether to avoid biases towards specific regions. To the extent that there are differences in safety regulations by state, however, this field could contain useful information. It may be unreasonable to compare two states which have different regulatory standards together. Including STATE in the model would allow for this to be taken into account. I will leave it up to the Union to decide if they want this included. Models were the STATE was included did result in an improve performance at the cost of interpretability.

A new feature was created based on the PRIMARY field. The original variable had 79 unique levels which is too many to use in a model. A new field called PRIMARY_BUCKET grouped this into simpler levels. This simplified the field down to five levels.

Model Selection and Validation

All models were trained on 80% of the data and validated against 20%. A random seed was set to ensure reproducibility. Models were compared based on four performance metrics. These were

- Mean Absolute Error (MAE). This penalizes outliers less than other risk measures which is good in this instance where I do not have time to examine every outlying case. Lower MAE is better.
- Root Mean Squared Error (RMSE). This penalizes outliers more but does a better job of converging. Lower RMSE is better.
- R^2. This is easy to interpret. Higher R^2 is better.
- Poisson Log Likelihood. This provides an indicator of how well the model fits the Poisson-model assumption. Higher likelihood is better.

Decision Tree

The decision tree was too complex initially. This was because there were variables with a large number of levels which made interpretation difficult. One weakness of decision trees is that because they consider all possible split values, variables which are continuous, or have a large number of values such as US_STATE or PRIMARY are more likely to be chosen. This results in the model overfitting.

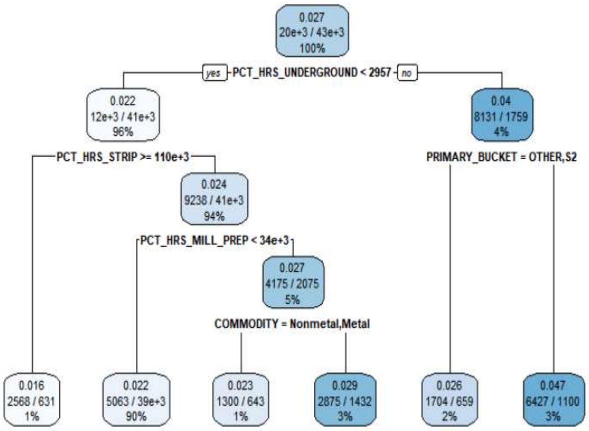
There were two ways that I corrected for this. 1) Increasing the complexity parameter (CP) and the minimum number of observations allowed in a node (min_bucket), and 2) by reducing the number of levels in the variables. I removed US_STATE and used a bucket for the PRIMARY field.

On the table below, Beryl's initial tree does have the highest log likelihood, but this is at the cost of the complexity of the model. If we were to penalize this with the number of coefficients in the model, such as with AIC or BIC, this would be worse. Model 2 is a much simpler version of model 1 and has a lower RMSE and MAE. The logLikelihood is slightly lower but this is still higher than the other trees, models 2 and 3. Model 2 was the final model selected.

Model	Description	RMSE	Rsquared	MAE	logLik
1	Beryl's Decision Tree	1.47	0.66	0.41	320.04
2	Higher min bucket with percentages converted to hours	1.35	0.66	0.40	268.40
3	Model 2 with a higher CP	1.35	0.66	0.41	256.20
4	Model 2 with a lower CP and higher min bucket	1.35	0.65	0.41	261.48

In the tree below, note that the “PCT_HRS_#” features are really in units of hours and not percentages. This was to save time. The result means that the injuries per 2000 hours for different mine types are

- **Low** (0.016) where workers spend less than 2957 hours underground and more than 110,000 hours in the strip
- **Low** (0.022) where workers spend less than 2957 hours underground, less than 110,000 hours in the strip, and less than 34,000 hours in the mill
- **Low** (0.023) where workers spend less than 2957 hours underground, less than 110,000 hours in the strip, more than 34,000 hours in the mill and whose commodity is metal or non-metal
- **Moderate** (0.029) where workers spend less than 2957 hours underground, less than 110,000 hours in the strip, more than 34,000 hours in the mill and whose commodity is not metal or non-metal
- **Moderate** (0.026) where workers spend more than 2957 hours underground, are in limestone or other type mines
- **High** (0.047) where workers spend more than 2957 hours underground, are in sand & gravel, coal, bituminous, or stone mines



GLM

To ensure that the model is not overfitting to a specific metric, in addition to the Poisson log likelihood that was provided I also compared the Root Mean Squared Error, R^2, and MAE. The final model selected is better across almost all of these metrics.

The modeling steps were interative. I started with Model 1, Beryl’s model, and fit 3 additional models.

There were a few issues with the first version of the GLM. I noticed that the algorithm was not converging, there were NAs in the coefficients of TYPE_OF_MINE, and the percentage variables had very large coefficients. In addition, there was no log transform applied to the continuous variables.

These large coefficients were fixed when the percentages were converted into hours by multiplying by total hours for the min, EMP_HRS_TOTAL.

I first attempted to fix the NA values by removing the records with Mills, which were the records causing the difficulty. This still resulted in errors and so I removed this field TYPE_OF_MINE entirely.

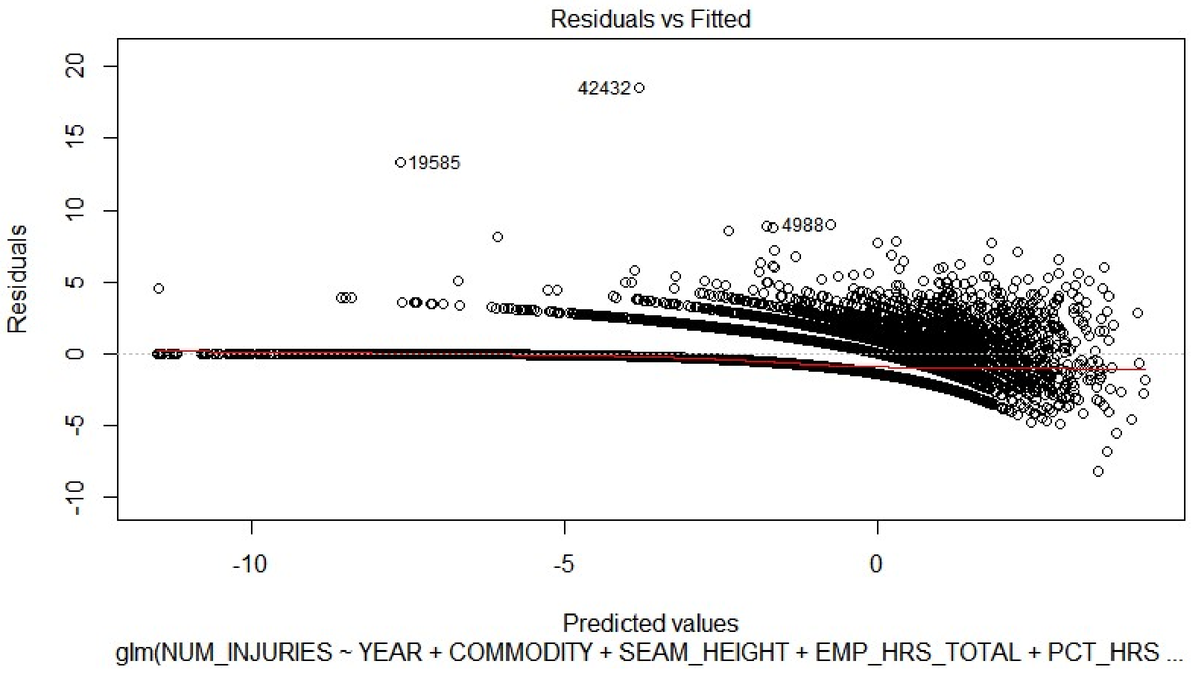
I applied a log+ 1 transformation to each of the hours by mine type fields. This led to a final log likelihood of 870 as compared to an initial -574, which I believe you can see is a huge improvement.

Model	Description	RMSE	Rsquared	MAE	logLik
1	Beryl's GLM	1.23	0.65	0.38	-574.53
2	GLM with percentages converted into hours	1.44	0.67	0.41	499.16
3	Model 2 with TYPE_OF_MINE removed	1.50	0.68	0.40	813.64
4	Model 3 with log transform of hours by all mine types	1.47	0.69	0.39	870.45

Finally, in order to make the results more interpretable I also fit a model excluding state. This had performance but is much more interpretable. This is the model which was selected for this draft. If the Union would like us to adjust by state this can easily be added back to the model.

```
     RMSE    Rsquared         MAE      logLik
1.6494775   0.5854409   0.4304482 484.3296945
```

One key assumption of linear models is that the residuals are uncorrelated and centered at 0. This is the graph of model 4 above. The residuals are centered at zero and are mostly random. There does appear to be some pattern which could indicate that an important predictor is omitted (perhaps we can ask the Union if there are other factors not being considered).



Findings

We have a set of rules which will predict the number of injuries that a mine has over 2000 hours. A recipe for this is as follows. The basic idea is to create a risk score which will tell how risky a particular mine is. The union can use this to create safety standards. This risk score will be the average number of injuries that the mine should experience in 2000 hours.

Start with 38.
subtract 0.02 times the year
add 0.19 if the commodity is stone
add 0.55 if the commodity is coal
subtract 0.03 if nonmetal
add 0.31 if metal
subtract 0.0002 times the seam height
subtract 17.5 times the log of total employee hours
add 17.520993 times the log of hours underground
add 17.520993 times the log of hours surface
add 17.520992 times the log of hours strip
add 17.520983 times the log of hours auger
add 17.520982 times the log of hours culm_bank
add 17.520992 times the log of hours dredge
add 17.520993 times the log of hours other_surface
add 17.520991 times the log of hours shop_yard
add 17.520993 times the log of hours mill_prep
add 17.520992 times the log of hours office

This will give you a risk score. To convert this to the number of injuries, raise 2.718 to this power.

Appendix

GLM coefficients and p-values

Term	Estimate	P-Value
(Intercept)	37.95	0.00
YEAR	0.0	0.00
COMMODITYStone	0.2	0.00
COMMODITYCoal	0.5	0.00
COMMODITYNonmetal	0.0	0.34
COMMODITYMetal	0.3	0.00
SEAM_HEIGHT	-0.0002	0.00
EMP_HRS_TOTAL	-17.520993	0.27
LOG_HRS_UNDERGROUND	17.520993	0.27
LOG_HRS_SURFACE	17.520993	0.27
LOG_HRS_STRIP	17.520992	0.27
LOG_HRS_AUGER	17.520983	0.27

LOG_HRS_CULM_BANK	17.520982	0.27
LOG_HRS_DREDGE	17.520992	0.27
LOG_HRS_OTHER_SURFACE	17.520993	0.27
LOG_HRS_SHOP_YARD	17.520991	0.27
LOG_HRS_MILL_PREP	17.520993	0.27
LOG_HRS_OFFICE	17.520992	0.27