Bryce Bowles
648 Business Data Analytics 901
11/24/2020

Alchemy Broker Executive Summary

Alchemy Insurance sought insight into evaluating and predicting broker performance based on historical data. Specifically, the team was tasked with segmentation analysis and predicting whether gross written premium will increase or decrease in the next year.

- Steps were taken to explore, visualize and describe five groups of brokers using principal component analysis
- Four predictive models were built, evaluated, and then tuned for prescriptive measures to analyze broker performance

Results: The top performing cluster was cluster 2. This cluster had higher gross written premiums for the past two years. A random forest model with a high AUC of 0.7321 was used to predict whether the 2020 Gross Written Premium will increase or decrease from 2019 with a misclassification rate of 35%. Important variables for prediction included gross written premium of the past two years and a total of the policy counts of the past three years. The csv file included with this report titled “rf_predictions.csv” contains the probability that the gross written premium will go up in 2020 for each broker id.

Recommendations:

The random forest model can be used to predict future brokers' performance, identify variables that drive results and impact the company to become more “data driven”.

Alchemy Broker Segmentation & Performance

Problem Introduction:

Alchemy Insurance is seeking ways to manage brokers more effectively and revise incentive compensation plans for the brokers as needed. To do this, Alchemy asked our team to evaluate and predict broker performance based on historical data. Specifically, they asked for a segmentation analysis and predictive model of whether gross written premium will increase or decrease in the next year.

Preprocessing:

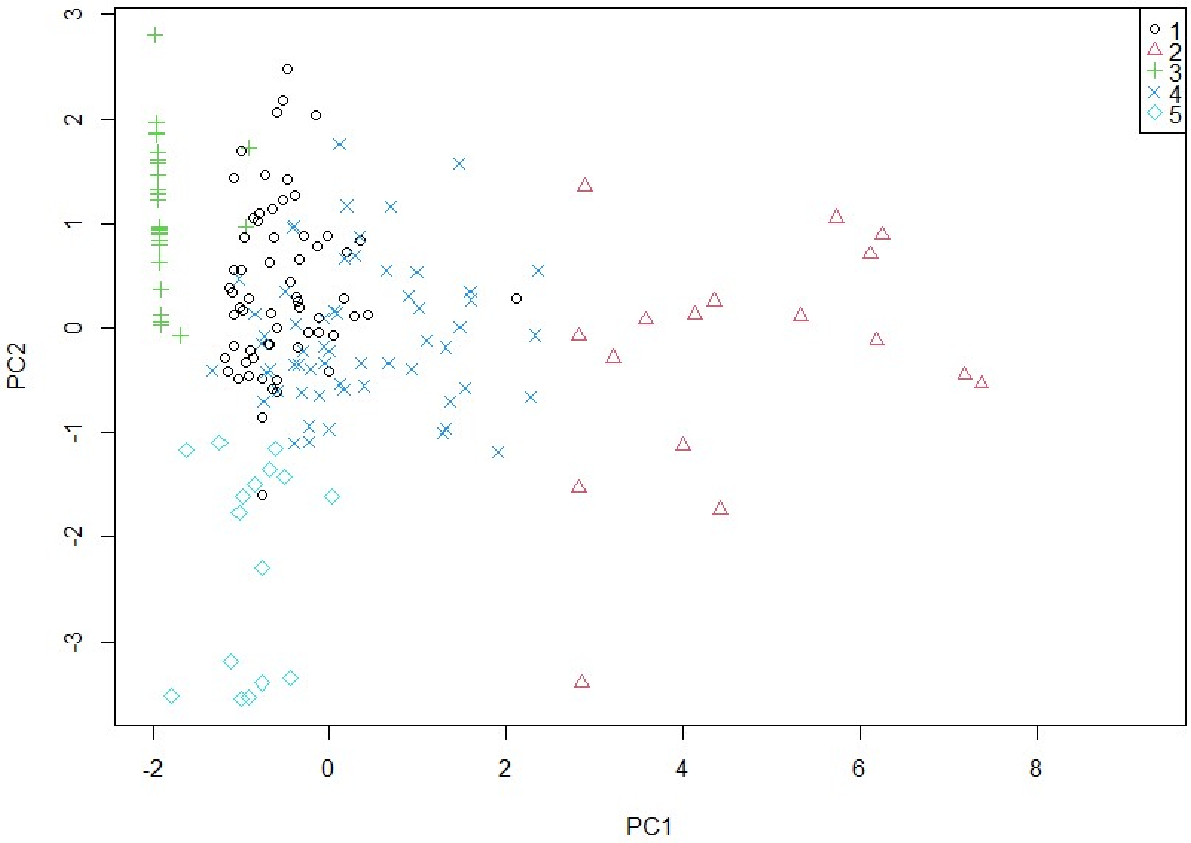
A data frame was created by selecting Submissions 2016 to 2018, Quote Count 2016 to 2018, Gross Written Premium 2016 to 2019, and Policy Count 2016 to 2018 from the broker dataset. The variables were renamed so that the model could be adapted to predict future years easily. Next, a variable was created called “up no”, which indicated whether the gross written premium had increased from 2018 to 2019. If it has increased it was assigned up, if not then no. Missing values were imputed with a 0. Then quote ratios for each year was calculated as the quote count of that year divided by the submissions. Also, a variable was created summing the policy counts of the three years named total policy. If any of the quote ratios were missing, then a 0 was assigned. If any quote ratios were infinity, then a 1 was assigned. Final variables were selected for the model and classification methods including gross written premium 2016-2018, quote ratios 2016-2018, and total policy. The data was partitioned, 75% training data and 25% test.

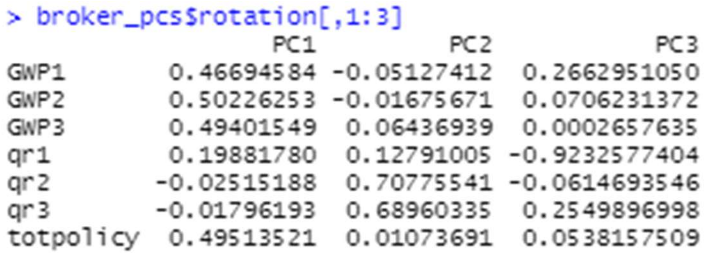
Broker Segmentation:

The predication data frame was used for clustering. This data frame's variables included: Gross Written Premium 2017-2019, Quote Ratios 2017-2019 and Total Policy Count of the three years. Summarizing the data after it was centered and scaled showed high max values and indicated that there would be outliers. This led to creating a cluster dendrogram to view hierarchical clustering by calculating pairwise distances between each observation. The graph is gradually right skewed with outliers that are higher than the averages toward the left of the graph. We can see broker 23 and others that are outliers and "far" from the majority of the dataset.

K-means clustering was then used to create 5 clusters. To visually evaluate the quality of the clusters and how well each point belongs to its cluster, a silhouette plot was used to view the principle component analysis (PCA) scores with an average silhouette width of 0.28. Cluster 2 and cluster 4 have brokers with negative coefficients, indicating poor cluster assignment. This means that there is variation in the data that we cannot see in the PCA plot that overlaps it from other clusters, however the other clusters seem to not have many negative coefficients.

A summary of the principle component scores was used to calculate a cumulative proportion and bar chart to show that the first two scores explain 77% of the variation in the original dataset.





```
> broker_pcs$rotation[,1:3]
                  PC1         PC2           PC3
GWP1       0.46694584 -0.05127412  0.2662951050
GWP2       0.50226253 -0.01675671  0.0706231372
GWP3       0.49401549  0.06436939  0.0002657635
qr1        0.19881780  0.12791005 -0.9232577404
qr2       -0.02515188  0.70775541 -0.0614693546
qr3       -0.01796193  0.68960335  0.2549896998
totpolicy  0.49513521  0.01073691  0.0538157509
```

Cluster 1, Black Circles: This cluster has positive values for PC2 and is centered around 0 for PC1. This indicates these brokers are likely to have high quote ratios in 2018 and 2019.

Cluster 2, Red Triangles: This cluster has positive values for PC1 and is centered around 0 for PC2. This indicates that these brokers are likely to have higher gross written premiums for 2018 and 2019 and higher total policy counts.

Cluster 3, Green Plusses: This cluster has negative values for PC1 and positive values for PC2 which indicates these brokers likely have large quote ratios in 2018 and 2019 and lower total policy count and gross written premiums in 2018 and 2019.

Cluster 4, Blue X's: This cluster is centered around 0 for both PC1 and PC2. This indicates that these brokers do not have extreme values for the corresponding variables relative to other brokers.

Cluster 5, Aqua Diamonds: This cluster has negative values for PC2 and negative values for PC1. This indicates that these borrowers likely have lower quote ratios for 2018 and 2019, lower gross written premiums in 2018 and 2019 and lower total policy counts.

Gross Written Premium Prediction:

Different classification methods were tested and analyzed including classification trees, logistic regression, random forests, and support vector machines. The results are shown below.

Misclassification Rates:

Classification Tree: 0.28

Logistic Regression: 0.35

Random Forests: 0.35

Support Vector Machine: 0.35

Area Under the Curve:

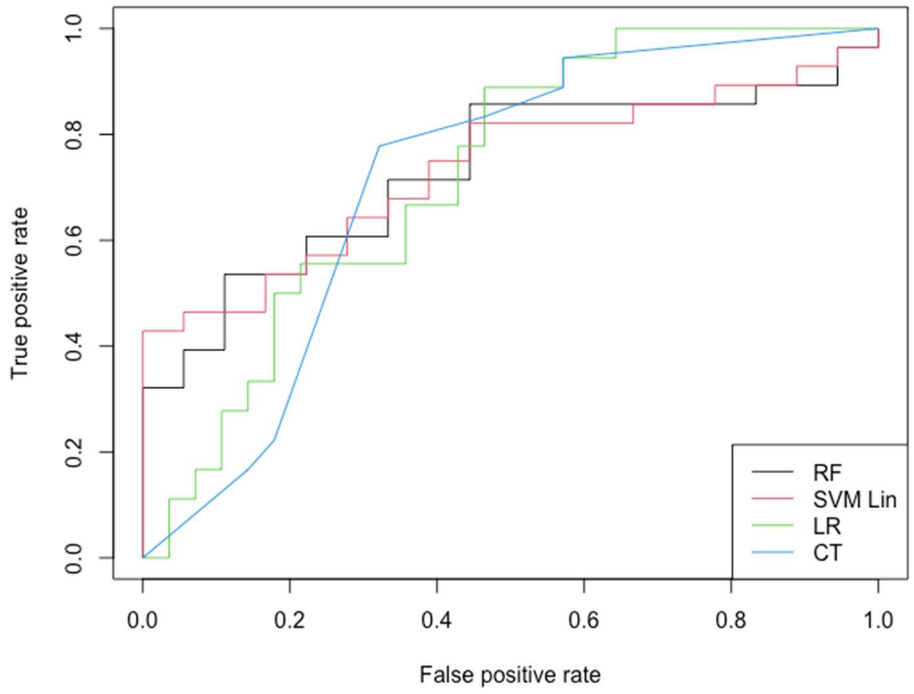
Classification Tree: 0.7143

Logistic Regression: 0.7242

Random Forests: 0.7321

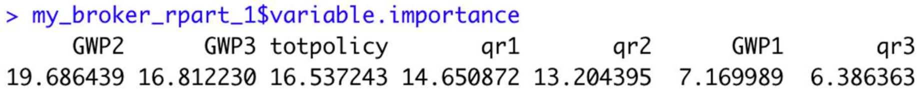
Support Vector Machine: 0.7341

ROC Curve:



Variable Importance:

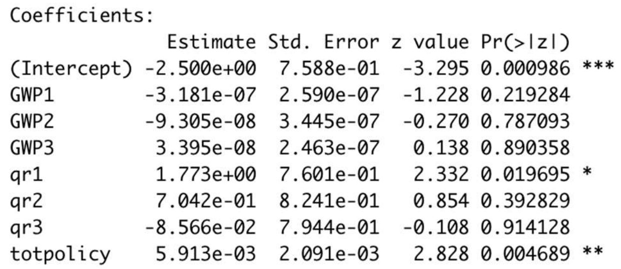
Classification Tree:



```
> my_broker_rpart_1$variable.importance
     GWP2      GWP3 totpolicy       qr1       qr2      GWP1       qr3 
19.686439 16.812230 16.537243 14.650872 13.204395  7.169989  6.386363 
```

Gross Written Premium 2018-2019 and total policy count of 2017-2019 are important variables for prediction.

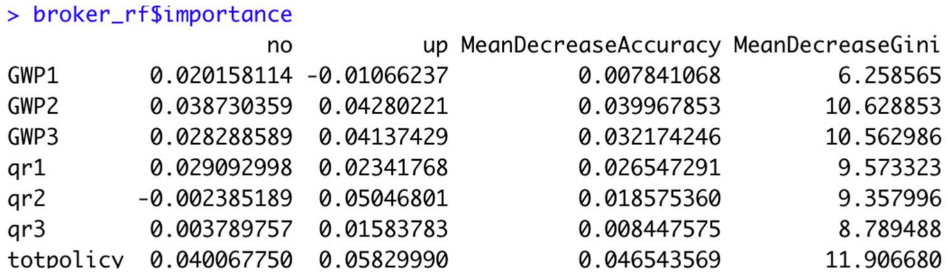
Logistic Regression:



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.500e+00  7.588e-01  -3.295 0.000986 ***
GWP1        -3.181e-07  2.590e-07  -1.228 0.219284
GWP2        -9.305e-08  3.445e-07  -0.270 0.787093
GWP3         3.395e-08  2.463e-07   0.138 0.890358
qr1          1.773e+00  7.601e-01   2.332 0.019695 *
qr2          7.042e-01  8.241e-01   0.854 0.392829
qr3         -8.566e-02  7.944e-01  -0.108 0.914128
totpolicy    5.913e-03  2.091e-03   2.828 0.004689 **
```

Total policy count of 2017-2019 and Quote ratio 2017 are important variables for prediction.

Random Forests:



```
> broker_rf$importance
                   no          up MeanDecreaseAccuracy MeanDecreaseGini
GWP1       0.020158114 -0.01066237          0.007841068         6.258565
GWP2       0.038730359  0.04280221          0.039967853        10.628853
GWP3       0.028288589  0.04137429          0.032174246        10.562986
qr1        0.029092998  0.02341768          0.026547291         9.573323
qr2       -0.002385189  0.05046801          0.018575360         9.357996
qr3        0.003789757  0.01583783          0.008447575         8.789488
totpolicy  0.040067750  0.05829990          0.046543569        11.906680
```

Total policy count of 2017-2019, gross written premium for 2018 and 2019 and quote ratio for 2017 are important variables for the random forests model.

Support Vector Machine: There is no easy way to assess the importance of predictors in SVM models.

Predicting Whether 2020 Gross Written Premium Will Increase:

The random forest model was adapted and applied to the predication data frame. Although the support vector machine model had a slightly higher area under the curve and the classification tree had a lower misclassification rate, the team chose to use a random forest model due to a high AUC while still being able to recognize variable importance. The team felt it was essential for Alchemy to understand the importance of which variables the model used for prediction to know what drives performance. The prediction data frame's variables included: Gross Written

Premium 2017-2019, Quote Ratios 2017-2019 and Total Policy Count of the three years. The csv file included with this report titled “rf_predictions.csv” contains the probability that the gross written premium will go up in 2020 for each broker id.

Conclusion:

Alchemy broker performance was evaluated by using historical data to explore, visualize and describe five groups of brokers using principal component analysis. Four different model types were then created to predict broker performance based on historical data. The random forest model performs better than random guessing with a high AUC of 0.7321 and a misclassification rate of 35%. This model can benefit Alchemy to become more “data driven” by predicting brokers' performance in future years while knowing which variables are important to drive broker performance.