STAT5003
COMPUTATIONAL STATISTICAL METHODS
ASSIGNMENT 2 FINAL REPORT
PROJECT GROUP 3
(TUTORED BY DR. KITTY LO)
NOVEMBER 6, 2019

lagr7305	**490617999**
mgup6878	**490537365**
rmen0735	**490462137**
grad0149	**490609961**

OVERVIEW

Life insurance businesses function given the prospect of risk associated with their clients. These insurance companies depend a lot on their systems to classify risks and produce different safety incentives, risk distributions, and protections against the losses (Abraham, 1985). They take into consideration an array of features related to the background of each prospective client, and based on the scores of each feature and the product/service they have opted for, the customers will be segmented accordingly, or in the case taken in this report, into one of the four risk levels (multi-class classification). This would, in turn, determine how favourably the insurance company would charge its customers and the premium they would avail (not in the scope of this project).

These features that are used for classification Coma broad range of aspects of the client, including BMI, employment information, medical history, insurance history and family background. Most of them are in turn subdivided (for instance, medical history is spread through six parameters - Medical_History_1-41). With over a hundred features to consider, there might be a need to implement feature selection and, possibly, PCA to reduce the dimensionality, before training the predictive models and evaluating their performances (Sigel, 2018).

DATASET DESCRIPTION

This dataset belongs to the insurance company ‘Prudential’, made available on Kaggle. The challenge here is to create predictive models to accurately classify the risks involved that could help the company quote better prices to its new and existing customers. The predicting power of the models created thereafter could help the company outline the importance of the data points in a more efficient way, enabling them to streamline the process of issuing their services (Prudential, 2015).

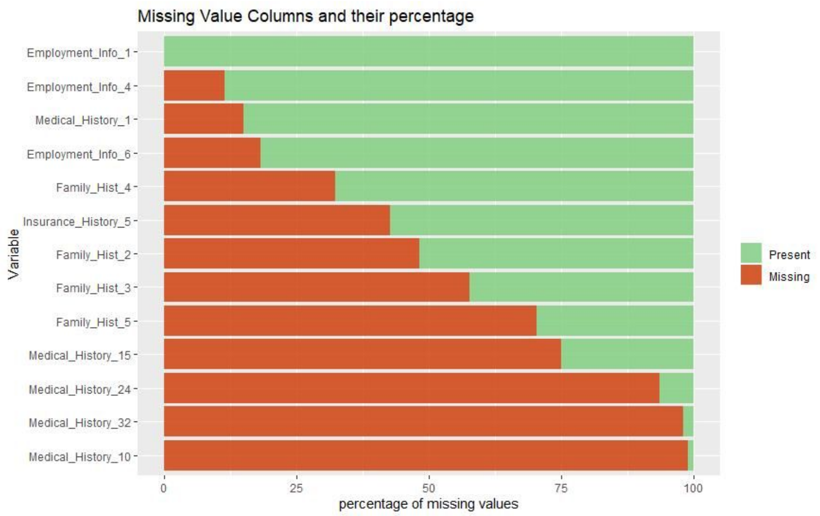
The dataset consists of 59,381 rows and 128 columns (with 13 primary fields, given below):

ID	**Unique identity for each application**
Product_Info_1-7 (7 columns)	Normalised variables related to the applied product
Ins_Age	Normalised applicant's age
Ht	Normalised applicant's height
Wt	Normalised applicant's weight
BMI	Normalised applicant's BMI
Employment_Info_1-6 (6 columns)	Normalised variables of applicant's employment history
InsuredInfo_1-6 (6 columns)	Normalised variables of applicant's information
Insurance_History_1-9 (9 columns)	Normalised variables of applicant's insurance history
Family_Hist_1-5 (5 columns)	Normalised variables of applicant's family history
Medical_Hist_1-41 (41 columns)	Normalised variables of applicant's medical history
Medical_Keyword_1-48 (48 columns)	Dummy variables related to a certain keyword's presence or absence in the application
Response	Target Variable related to application's final decision of risk classification

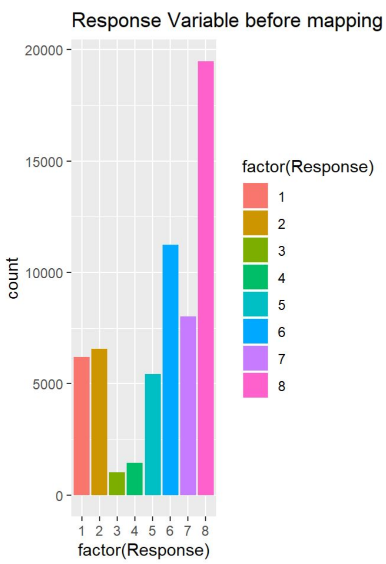
(Prudential, 2015)

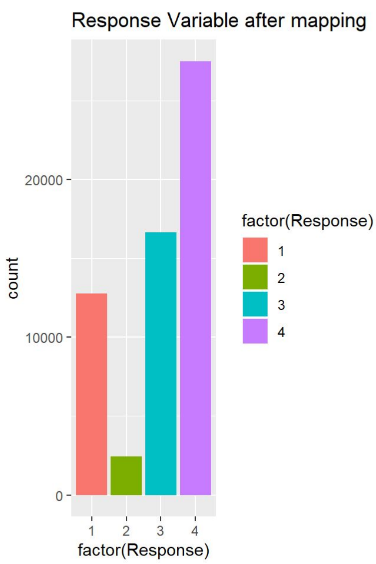
The dataset comprises of a mixture of data types - categorical (60 fields), continuous (13 fields) and discrete (5 fields). The medical keywords are dummy variables (48 fields).

The response contains 8 risk levels, which are then mapped into 4 risk levels by combining every two consecutive levels into one.

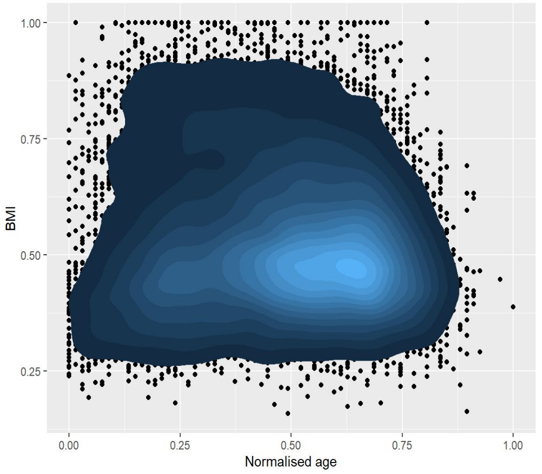


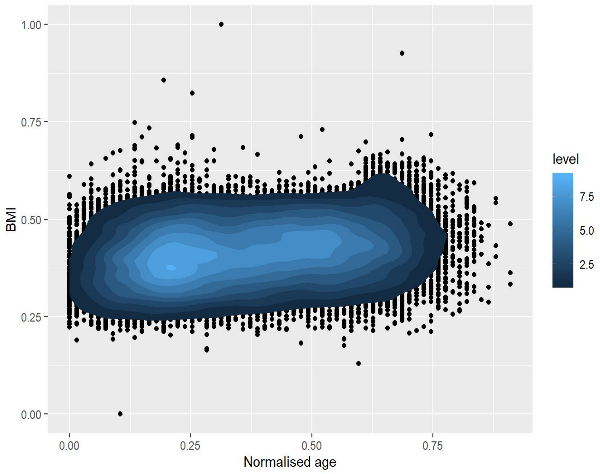
As is represented by the bar graph above, the dataset is imperfect due to the presence of missing values, with some fields like 'Medical_History_32' and 'Medical_History_10' having more than 90% of the values were missing. Such fields would require pre-processing.



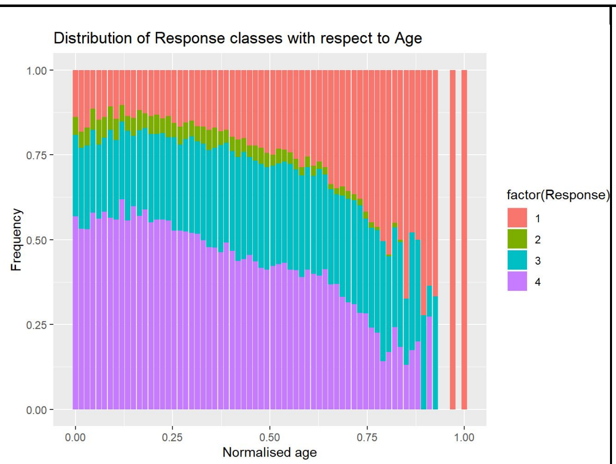


The response variables are not equally distributed in the train data, with risk levels '3' and '4' (Risk level '2' after mapping) only representing 1.7% and 2.4% of the whole 'response' column respectively. This could affect the predictive power of the models when it comes to predicting such levels on the test data (imbalanced classes).





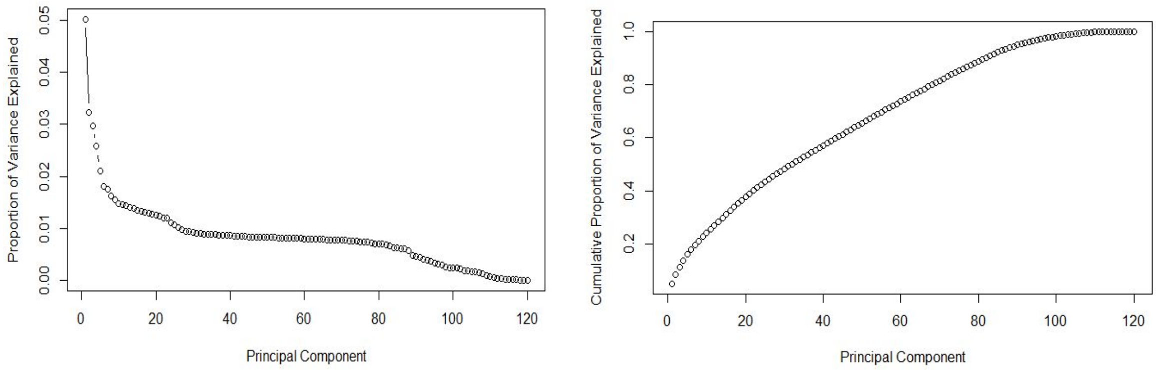
The plot on the left represents how BMI varies with age for level 1 (high risk) clients, and the other plot represents the same for level 4 (low risk) clients. Although a substantial number of clients have their BMIs closer to 0.5, level 1 clients have a more variance in the distribution for all ages (0.25 to 0.9), and are hence classified as high risk, while level 4 clients have a smaller variance (0.25 to 0.5), and hence poses less risk to the company. Note that higher concentrations of data points are represented by lighter shades in both the plots.



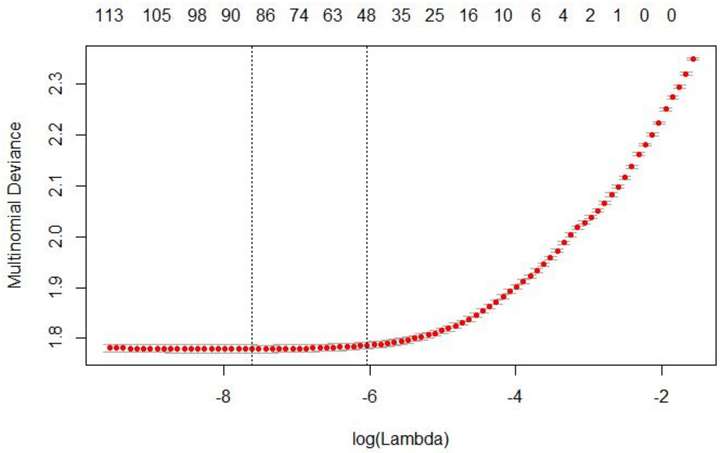
The plot on the left indicates how the clients are classified into the respective levels with respect to the frequencies of different ages. Youngsters are considered to be less risky (more blue and less red), whereas older clients are more risky (less blue and more red). Hence, they are likely to be classified as a level 1 or 2 while the former as a level 3 or 4.

FEATURE ENGINEERING AND DIMENSIONALITY REDUCTION

The scope of feature engineering is limited on this dataset as most of the fields are either masked or normalised at source. However, we were able to make use of the count of 48 medical keyword features, a set of binary valued features relating to the presence/absence of a medical keyword being associated with each client (row). This is to explore whether a client who is associated with a higher number of medical keywords is more likely to be classified as “high” risk.



For reducing the dimensions, PCA was applied. It can be seen that the first 100 principal components cumulatively captures 97% of the variance. However, PCA was not found to be useful, with only a marginal improvement over what would be otherwise, and was thus disregarded.

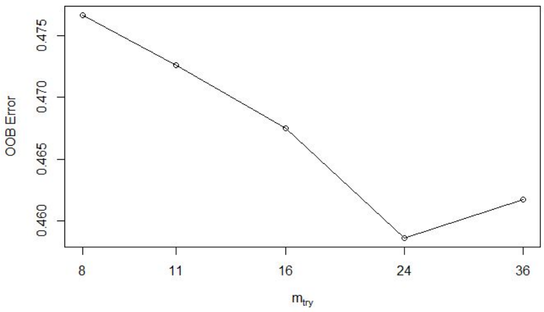
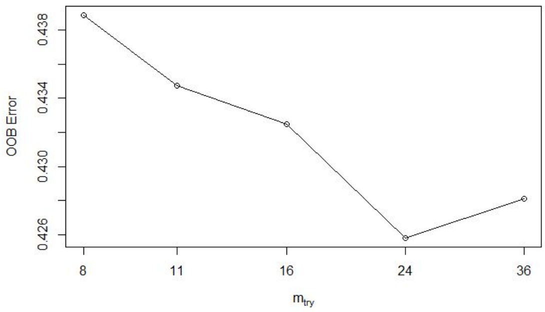
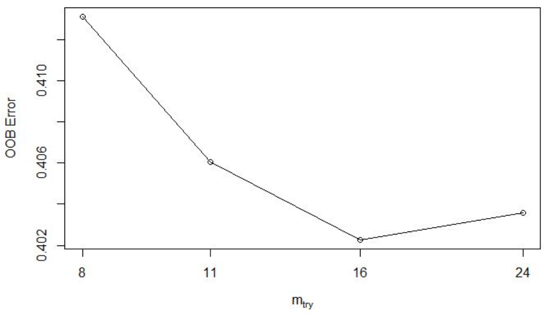
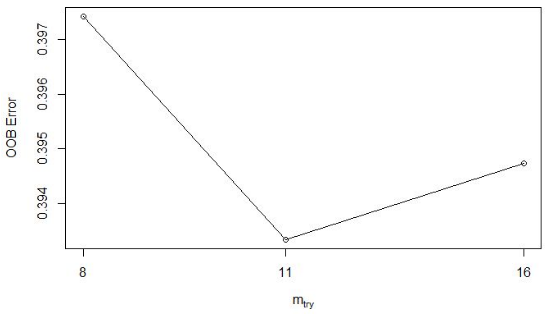


Lasso regression was used to find the features having non-zero coefficients. Although there was no improvement in the overall performance of logistic regression model, the cost of computation was reduced by taking into account only 74 of the otherwise 123 features.

HYPER-PARAMETER TUNING ON CLASSIFICATION MODELS

Random Forest classifier

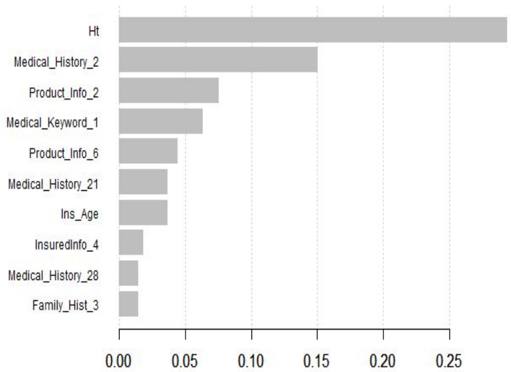
The dataset has 122 features. By default, roughly 11 features would be used at each split for making the tree. *Maxnodes* has also been restricted to a maximum of 20 in order to reduce the cost of computation. The *number of trees* used is 5,10,15 and 20. Other than manual tuning, *mtry* for each of these has been tuned using *tuneRF()* to reduce the out-of-bag error.

 Number of trees = 5	 Number of trees = 10
 Number of trees = 15	 Number of trees = 20

For instance, for 15 trees, the value of mtry that minimizes OOB error is 16.

XGBoost

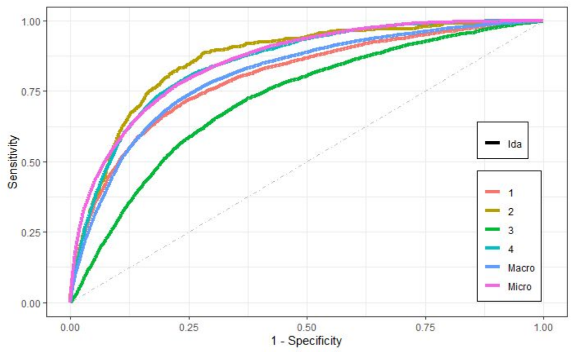
By setting the maximum depth to 2 and keeping *nrounds (300,500,1000)* constant, the parameters *gamma (0,0.50,1)* and *eta (0.4,0.6,0.8,1)* were tuned using GridSearch.

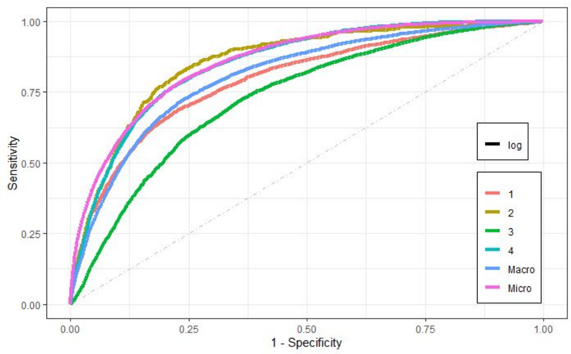
	The graph on the left is a representation of the information gain for the top 10 features that were found to be influential for making key decisions by the XGBoost classifier, after hyper-parameter tuning. It is worth noting that height and age are among these features.

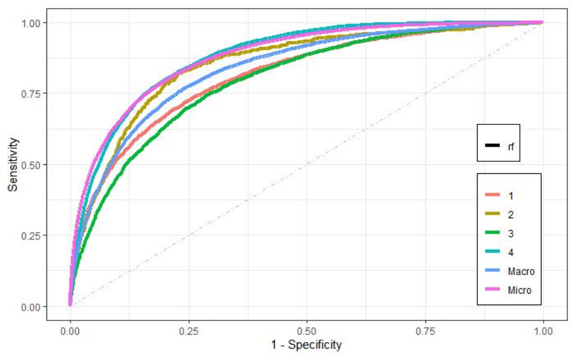
Lasso regression was performed and was used to model logistic regression. Although it reduced the number of features to 74 and reduced the computation, there was no significant improvement in the metrics.

EVALUATION

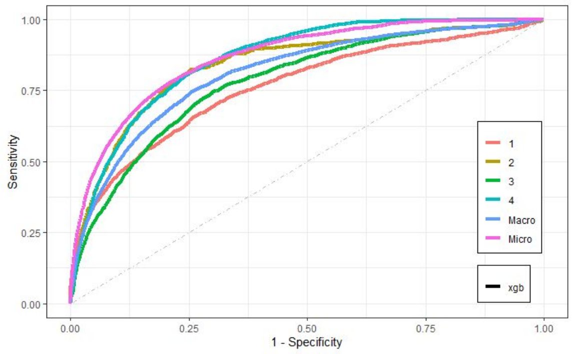
The four models have been compared using AUC-ROC scores for each of the four levels, the average AUC score, and the F1 scores (Belik and Gauher, 2016).

<table>
<tr><td>ROC curve</td><td>Per-class AUC scores</td><td>Average AUC</td><td>F1 score</td></tr>
<tr><td>


LDA</td><td>
<table>
<tr><td>1</td><td>0.7931</td></tr>
<tr><td>2</td><td>0.8711</td></tr>
<tr><td>3</td><td>0.7300</td></tr>
<tr><td>4</td><td>0.8498</td></tr>
</table>
</td><td>0.811</td><td>0.6214</td></tr>
<tr><td>


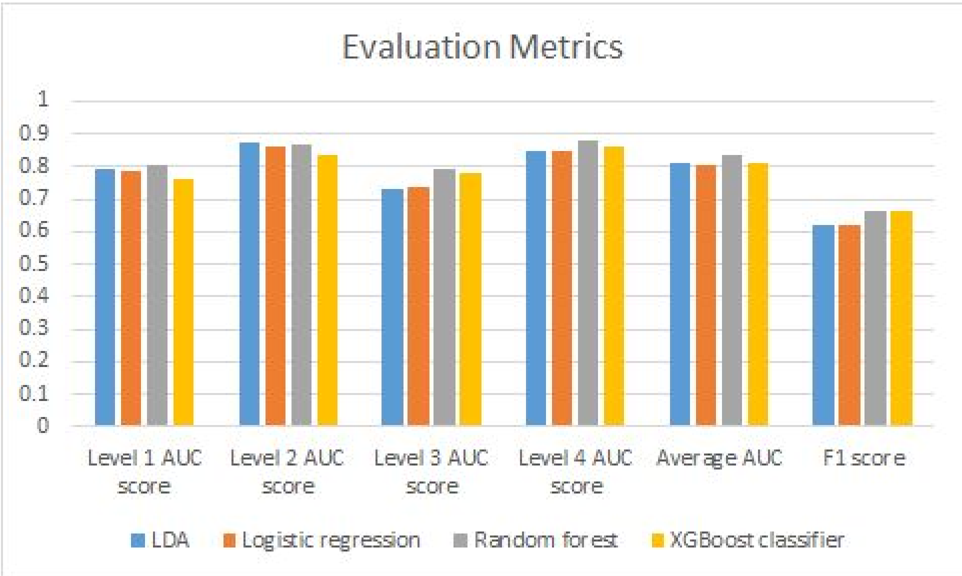
Logistic</td><td>
<table>
<tr><td>1</td><td>0.7870</td></tr>
<tr><td>2</td><td>0.8597</td></tr>
<tr><td>3</td><td>0.7387</td></tr>
<tr><td>4</td><td>0.8501</td></tr>
</table>
</td><td>0.808</td><td>0.6209</td></tr>
<tr><td>


Random forest</td><td>
<table>
<tr><td>1</td><td>0.8089</td></tr>
<tr><td>2</td><td>0.8653</td></tr>
<tr><td>3</td><td>0.7958</td></tr>
<tr><td>4</td><td>0.8809</td></tr>
</table>
</td><td>0.837</td><td>0.6648</td></tr>
</table>

<table>
<tr>
<td>

XGBoost class</td>
<td><table><tr><td>1</td><td>0.7609</td></tr><tr><td>2</td><td>0.8383</td></tr><tr><td>3</td><td>0.7814</td></tr><tr><td>4</td><td>0.8609</td></tr></table></td>
<td>0.810</td>
<td>0.6624</td>
</tr>
</table>

Random forest classifier performs the best overall with the highest AUC-ROC score (Horton, 2016). This was then followed by XGBoost. All the four models are relatively better at classifying level 2 and level 4 clients (higher AUC scores). Level 4 showed high sensitivity, probably because of a majority of the observations in the training set belonging to that level.



The above image represents the performance recorded by the evaluation metrics for each of the models.

CONCLUSION

The insights gained from EDA indicate that the insurance policies for youngsters are less risky for the companies as against to the aged people. Also, the BMI for the less risky clients lie close to the general BMI while it has a high variance in case of more risky clients. The classification models applied on the problem gave better performance for the Random Forest and xGBoost as compared to Logistic Regression and LDA. Also, the dataset was normalised at the source and during data pre-processing the feature selection using Lasso Regression and Feature scaling using min-max scaler were not found to be of much help as they were unable to enhance the performance of the models.

REFERENCES

1. Abraham, K S 1985, Efficiency and Fairness in Insurance Risk Classification, Virginia Law Review, vol. 71, no. 3, pp. 403-451.
2. Prudential 2015, Prudential Life Insurance Assessment, Kaggle, viewed 14 September 2019, available at: https://www.kaggle.com/c/prudential-life-insurance-assessment/overview
3. Sigel, Z 2018, What are life insurance classifications? Policygenius, viewed 21 September 2019, available at: https://www.policygenius.com/life-insurance/what-are-life-insurance-classifications/#how-to-get-a-better-life-insurance-classification
4. Bleik, S and Gauher, S 2016, Computing Classification Evaluation Metrics in R, Revolutions, viewed 30 October 2019, available at: https://blog.revolutionanalytics.com/2016/03/com_class_eval_metrics_r.html
5. Horton, B 2016, Calculating AUC: the area under a ROC curve, R-bloggers, viewed 30 October 2019, available at: https://www.r-bloggers.com/calculating-auc-the-area-under-a-roc-curve/