Data Mining and Decision Systems
600092
Assigned Coursework Report

Student ID: 201707408
Date: 07 October 2019

Methodology - CRISP-DM

Business understanding

Understanding why the model is being created is a key part of the CRISP DM methodology as without a clear understanding of the goals the data scientist may remove important features or may train the wrong model entirely. Therefore, to save time and resources, the data scientist must fully understand the business requirements before the continue.

The data consists of visits to a hospital due to cardiovascular complications. The model being developed will be used to identify patents that are at risk, and therefore there are significant ramifications if the model is malformed therefore false negatives must be avoided when selecting a model. Data protection is also a significant concern however as we only have access to numerical identifiers and therefore identifying individuals would be extremely difficult. The model itself will be used to classify patients to risk or no risk depending on a variety of features, this however may have to be overseen by a trained medical professional due to regulations in the industry. It may also put patients at ease knowing a professional has looks at their case and decided on the best course of action with the model supporting their actions. This would possibly also protect them from legal issues if they relied fully on the trained model, therefore the main goal of the model is to assist medical professionals in identifying at risk patients allowing them to deal with patients more efficiently and confidently.

Data understanding

Understanding the data will allow the data scientist to find invalid values for the next stage of CRISP DM, without this data could be misunderstood and be modified incorrectly drastically affecting the accuracy of the final model. Visualization though not a part of CRISP DM allows the data scientist to spot key features in the data set, as well as redundant ones reducing the number of features which not only reduces complexity but could also reduce training times.

Reading the data description gave the data scientist insight into the domains of the columns and suggested. Comments were also provided specifying what each column means along with non-clinical descriptions. This allowed the data scientist to research how certain features could affect risk. Certain medical were harder to decipher such as ischemic and contralateral. Contralateral means opposite side of the body and Ischemic is defined as lack of oxygen leading to necrosis or cell death. Therefore, this means IPSI and Contra are the percentage of oxygen lacking lesions for each side of the body, and the closer they are to 100% the more risk that is involved. This would be verified with a medical professional or the data owner however this is one of the issues working with legacy data. Along with the lack of contextual data on where the data was does not cause issues with data understanding but may provide an insight into why model may have behaved in a certain way.

Looking at the data types and using the describe function in pandas allows the data scientist to get a good overview. Session ID did not seem to be provided and random seemed to have duplicates which did not match the data description which implied that either the data description or data is wrong. The unique values per column showed that many columns contained invalid values as many columns that are of Boolean domain. Other sets as is the case for the Indication feature are categoric but are within the cardiac domain with a limited set of values.

Data prep

Data preparation is required as machines do not understand human words such as ‘yes’ and ‘no’ well and this has a detrimental effect on model accuracy (Jaitley, 2018). And therefore, encoding the values to numerical data not only helps with visualization but model accuracy Due to the data containing invalid or incomplete values, cleaning was required. Removing or imputing these data values would allow for more consistent and accurate visualizing and models.
Using the data description as a guide to ensure data integrity, the data scientist checked the data types as this would aid with finding the domain of the columns, allowing him/her to spot and invalid values.
Firstly, the entire random column seemed to be to contain duplicates, therefore having a closer look at these data points suggested that these we not identical patents as they had varying features even if random was identical.

Setting an Index would be useful to call individual records so the data scientist can pull records by ID and could review them individually. ID was chosen as it was already unique and therefore could also be integrated into other legacy systems when deployed.

Contra was detected as an object, this made numerical analysis of the feature difficult. Forcing the column to numeric would make invalid values into NaN values which could be then be imputed or dropped.

Checking each feature’s unique values showed ‘Unknown’ in the label column, as the model being trained would rely on the label column keeping these records would be meaningless. ASX had 2 variants and with further research it was found there was only one ASX and therefore the capitalization was fixed to be inline with the other indication values.

Listing all the null values allowed the data analyst to decide whether to impute or drop the records containing null values. However, there were only 20 null records in total which made up only 1.32% of all the data and therefore removing them would not be detrimental to the models.
Random and ID did not affect the label and therefore was dropped as this would just interferer with the model training process and add to domain complexity unnecessarily.

Visualization

This not a part of crisp - DM however visualization would be helpful in identifying key features that affect risk and spot any outliers that may significantly affect model training. However, since Contra and IPSI were the only numeric features only these could be theology visualized.

Encoding binary values to numbers allowed the models to be trained more effectively as machine understand 1 and 0 better than yes or no. This was due using label encoder as this would replace values as opposed to creating new columns like get_dummies. Converting these to numeric allowed the data scientist to look at the correlation heat map and find that indication had a very weak correlation. Therefore, visualizing this feature was necessary to examine if the data was relevant to risk. There was a strong relation that TIA was an indication you are likely to be not risk and therefore a valid indicator of risk inline with the other features so it was decided it would not be removed.

Models

Logistic Regression

Logistic regression is used for classification models and therefore was chosen over linear regression. Being one of the simpler models allows the data scientist to quickly spot error in the data and model train process. Initially contra was chosen as this showed a strong correlation the visualization section. This had an accuracy of 84% however as all the features had a correlation to label adding the other features would increase accuracy further however would limit our ability to visualize the data.

Combining the features improved accuracy dramatically however due to the indication being grater than 1 the class weight had to be balanced as there wasn’t a strong correlation between label and indication. The solver by default is lbfgs however this is used for multi class classification, therefore this was substituted by liblinear, as this is designed for smaller datasets and binary classification. This did cause convergence issues therefore max iterations were increased to 300 iterations. The penalty value by default it l2, however l1 seem to have performed better as each feature was not equal and l1 has built in feature selection making it more robust than l2. This resulted in much higher accuracy, however overfitting was a concern and therefore the data analyst did cross validation with the model, this also resulted in the same and therefore we can conclude that overfitting did not occur and therefore the model has formed correctly.

Decision Tree

A decision tree is a very transparent model where it can solve both regression and classification problems and the data does not have to be normalized however normalization of the data tends to produce better accuracy and therefore it was generated with normalized data. The transparency of the decision tree allows the data scientist to visualize key features and relations between the data giving further insight into how the data is correlated. A max depth could have been set to prevent overfitting as decision tree did not have regulation penalty like logistic regression, however running cross validation on the model suggested it was not over fitting and produced the one of the most accurate models at 99% with cross validation suggesting the model had 99% accuracy. However, issues are caused if retrained with different data as this may produce an entirely different structure. This has significant ramifications when deployed as the model may be retrained with new data.

Neural net

The data scientist used a Multilayer perceptron classifier as this was a classification problem. It has many parameters and therefore has a lot of flexibility compared to other models and therefore could be tuned for the best accuracy. The sdg solver allowed for the data scientist to control the learning rate as this was a hyperparameter. A constant learning rate performed poorly as each feature did not have equal weighting that contributed to the result and therefore an adaptive learning rate was the best fit however this was outperformed by other solvers. The default solver was adam however due to the small size of the dataset, lbfgs was a better fit and the batch size also had to be lowered to 10 this boosted accuracy by 4% to 98% and cross validation confirming a higher than average variance. A random state had to be set as the training results would otherwise be inconsistent.

Random Forest

Random forest is an ensemble learning method for classification as this creates multiple decision tree models and averages them for a theoretically better score. As the decision tree was the best scoring model replicating this would theoretically increase accuracy. Increasing the number of estimators did not increase accuracy of the model and therefore was kept to 50 to reduce fit and prediction time, this was likely due to the size of the data set as random forest is designed to handle large amounts of data.

Note, mean squared error was not used as binary classification is a bad use case for it, due to it relying on the underlying data being from a normal distribution (a bell-shaped curve).

Results

All models use normalized datasets

Logistic Reg (n = 375)	Predicted No Risk	Predicted Risk	
Actual No Risk	TN=236	FP=7	243
Actual Risk	FN=4	TP=128	132
	240	135	

Decision tree (n = 375)	Predicted No Risk	Predicted Risk	
Actual No Risk	TN=241	FP=2	243
Actual Risk	FN=2	TP=130	132
	243	132	

Neural net (n = 375)	Predicted No Risk	Predicted Risk	
Actual No Risk	TN=240	FP=3	243
Actual Risk	FN=3	TP=129	132
	243	132	

Random forest (n = 375)	Predicted No Risk	Predicted Risk	
Actual No Risk	TN=239	FP=4	243
Actual Risk	FN=2	TP=130	132
	241	134	

Model	Recall	Precision	Score	F1 Score
Logistic Regression	0.9697	0.9481	0.9707	0.9588
Decision tree	0.9848	0.9848	0.9893	0.9848
Neural net	0.9773	0.9773	0.984	0.9773
Random Forest	0.9848	0.9772	0.984	0.9772

Model (Cross val average)	Fit time	Predicted time	Score	Variance
Logistic Regression	0.0253	0.0010	0.9733	0.00011
Binary tree	0.0030	0.0010	0.9893	$6.2_{x10}{}^{-6}$
Neural net	2.806	0.0024	0.9653	0.0014
Random Forest	0.0656	0.005	0.9888	$6.48_{x10}{}^{-5}$

Evaluation & Discussion

Evaluation of the models were done using a variety of metrics but Recall and accuracy were the most important as getting a false negative also known as a type II error would have significant consequences for the patient if not detected. A patient who is false positive on the other hand would receive additional tests to confirm their condition or at worst receive unnecessary treatment.

The logistic regression model had an accuracy of 97% which is very good however relative to other models it had the lowest score, recall, precision and f1 score. As recall suggested with 4 false negatives it was the worst model and therefore has the most potential to label patients at not at risk when are. Along with the highest number of false positives this is the worst model out of the 4. Variance for this model however is the lowest as this is not a very adaptive model and therefore should be robust against new data.

The decision tree performed surprisingly well considering no parameters we put in place to tune it to the particular data set. This expected this model to overfit however cross validation confirmed it predicted values correctly consistently. Looking closer at the tree itself, you can see the model already has a good idea of whether it is risk or no risk at the root node with a GINI index of less than 0.5. being the model with the highest accuracy and being on par with the random forest model makes this a strong contender. However, if retrained with a slightly different data set the entire structure could change which makes it harder to predict when deployed. Due to the nature of the data, a low variance is preferred, as medical must be consistent and accurate as possible.

The neural net was a computationally intensive compared to other models and had virtually identical metrics to the random forest model, performing only marginally worse than the random forest model. The recall and precision were affected accordingly with 3 false negatives and 3 false positives. However, with further tuning of the various hyperparameters, the model is very could possibly be improved, and modified to fit future datasets extremely well.

The Random forest performed extremely well and matched the decision tree in recall with 2 false negatives in line with the higher recall value. It did score slightly lower however random forest is more versatile than a single decision tree as random forest tend to have lower variance.

With cardiovascular disease being the biggest cause of death in the world (17.9 million deaths globally), 1500 patients are not enough for an accurate and consistent model to be created (WHO, 2019). The are a variety of environmental factors that could also affect the result however this introduces the 'curse of dimensionality' where the data could have too many features.

As a result, the best model would have to be the random forest classifier as it would be more consistent when trained with similar data sources and therefore could have more accurate results with new unseen data. Along with the added benefit it could be paralleled and therefore allow fast training times with larger datasets, allowing good scalability for actual deployment.

Outliers were also a concern as it could have affected model training, and this would be one of the reasons why the models never reached 100% along with the risk of overfitting.

The CRISP-DM methodology being the most popular cross industry standard functioned well during the assignment, with only real issues encountered was that some data could not be visualized as they had to be prepared into numerical values. Visualization in itself was not mentioned in crisp DM but as part of data understanding the data analyst used visualization to get a clearer picture of the data. Sk-learn's neural net is not used very much in industry as TensorFlow, a alternative framework is more feature rich and has more activation functions such as SoftMax, however this is out of scope for the assignment.

References

Jaitley, U., 2018. *Why Data Normalization is necessary for Machine Learning models.* [Online]
Available at: https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029
[Accessed 05 12 2019].
scikit-learn developers, 2019. *SciKit-learn-API-Ref.* [Online]
Available at: https://scikit-learn.org/stable/modules/classes.html
[Accessed 30 08 2019].
WHO, 2019. *Cardiovascular Diseases.* [Online]
Available at: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1
[Accessed 10 12 2019].