Predictive Purchase Behaviour Modelling - Motor Policies

Axil Sudra

Hastings Direct

March 3, 2020

Outline and Exploration

Objective

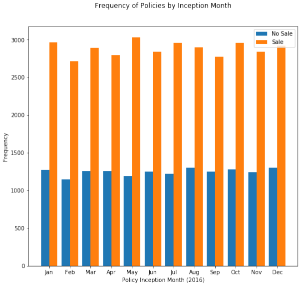
- Predicting customer behaviour in the personal lines Insurance market has always been important because of various benefits that can be gained (e.g. cost-cutting on marketing to unlikely customers).
- Hence the objective of this presentation is to explore and compare three different machine learning models that can predict whether or not a customer is likely to purchase a motor policy from Hastings Direct.
- Machine learning models used in this presentation: Random Forest, Support Vector Machine (SVM), and Multi-layer Perceptron (MLP).

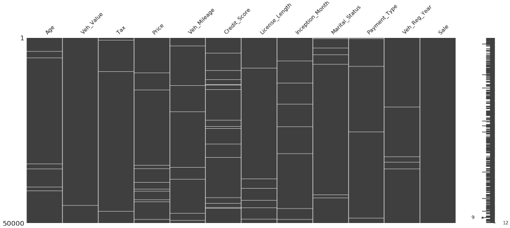
Dataset Description

- The dataset used to analyze and evaluate the three models is based on 'motor' quote data.
- This dataset consisted of $50,000$ observations with 11 predictor columns of which 7 are numerical, 3 are categorical, and 1 is a date.
- It also contained the class column 'Sale' to denote whether a customer did or did not purchase a motor policy from their existing quote. Note that this was the attribute to be predicted. The class distribution of the dataset was imbalanced with 30% of observations not purchasing a policy and 70% of observations purchasing a policy. Note that the Synthetic Minority Over-Sampling Technique (SMOTE) algorithm was used to address the class imbalance.
- Furthermore, the dataset contained $5,496$ missing values.

Data Wrangling and Predictor (Feature) Engineering

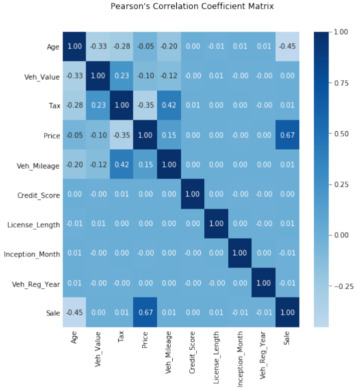
- Machine learning models cannot interpret data other than numerical data; therefore, the numeric month of the 'Date' predictor was extracted and stored in a predictor called 'Inception_Month'. Categorical predictors were also label encoded, and thus were represented numerically.
- Missing values were addressed through a imputation method known as Multiple Imputation by Chained Equations (MICE) to avoid discarding data (i.e. removing observations with missing values).

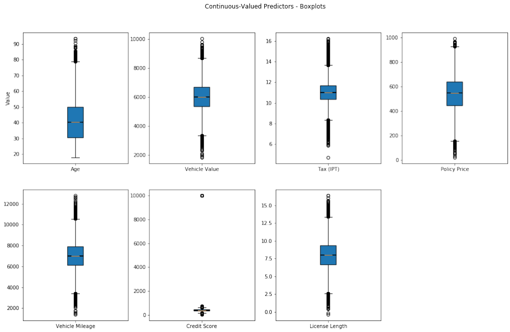




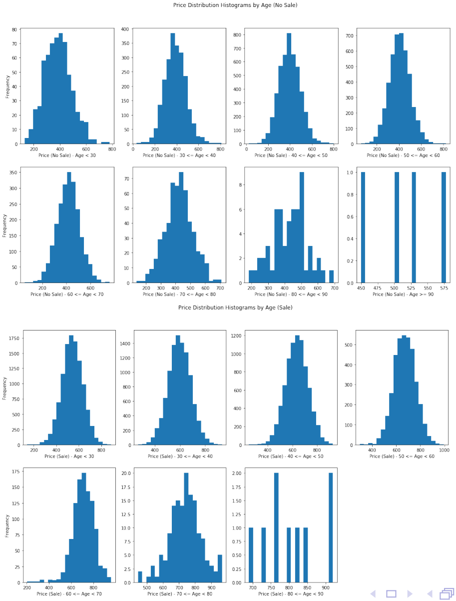
Exploratory Data Analysis

- A correlation heat-map was created to identify any significant correlation between predictors and class. Furthermore, ‘extreme’ outliers were identified via box plots and set to null if necessary (these were estimated by MICE).





Exploratory Data Analysis (Cont.)



Random Forest Description

- A model consisting of an ensemble of decision trees based on bootstrap aggregation (known as bagging).
- Each decision tree in the model is generated from a bootstrapped training dataset using random predictor selection; this is a method that selects a random subset of predictors at each node in the tree to reduce correlation between trees in the overall forest.
- Information gain can help improve the node splitting choice towards smaller trees.
- For classification tasks the model uses a majority vote system for class prediction - an instance is sent down every tree in the model and the most common class across all trees is assigned to that instance.

Support Vector Machine Description

- A model that attempts to identify a hyperplane in an N-dimensional space that distinctly classifies data observations to the correct class; note that N usually is given by the number of predictors in a given dataset.
- The hyperplane can be described as a decision boundary where the distance of separation between data observations from one class and another class is maximized.
- Note that 'support vectors' are the data observations that lie closest to the decision boundary (hyperplane).
- Data does not have to be linearly separable for SVMs to work - kernel trick!

Multi-layer Perceptron Description

- A type of artificial neural network constructed of an input layer, hidden layers, and an output layer. Note that each layer (excluding the input layer) consists of neurons that are formed of a linear summation function and a nonlinear activation function.
- Each layer in the network is fully connected via adjustable synaptic weights.
- A popular method for training MLPs is the back-propagation algorithm; a method known as stochastic gradient descent is used to optimize the synaptic weight values with respect to some loss function (e.g. binary cross entropy).
- Training consists of two phases; the forward phase and the backward phase. These phases are repeated until some loss function has been minimized or a certain amount of passes through the network have been completed (epochs).

Training and Evaluation Methodology

- A 75%/25% train/test split was used.
- A random search was used to source the 'best' set of hyper-parameters for each model; 5-fold cross validation was also implemented in the training process to mitigate the risk of overfitting to the training data.
- Evaluation methodology for each model included a confusion matrix from which the accuracy, precision, recall, and f1-score were calculated. A roc curve was also plotted and its corresponding auc score was retrieved.
- A number of various hyper-parameters were chosen to be tuned for each model; these are give on the next slide.

Training and Evaluation Methodology (Cont.)

Random Forest Classifier
• Number of trees = range [100, 1000]
• Max features (predictors) = [‘auto’, ‘sqrt’, ‘log2’]
• Max depth of trees = range [3, 10]
• Split criterion = [‘gini’, ‘entropy’]

Support Vector Classifier
• C = range [0.0001, 1]
• Gamma = range [0.0001, 1]
• Kernel = [‘rbf’, ‘linear’, ‘sigmoid’ ‘poly’]

Multi-layer Perceptron
• Number of hidden layers = range [1, 3]
• Number of hidden neurons = range [50, 300]
• Activation function = [‘logistic’, ‘tanh’, ‘relu’]
• Learning optimizer = [‘lbfgs’, ‘sgd’, ‘adam’]
• Learning rate = [‘constant’, ‘invscaling’, ‘adaptive’]

Model Building and Results

‘Best’ Model Hyper-parameters

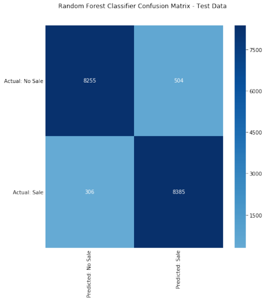
Random Forest Classifier
• Number of trees = 634
• Max features (predictors) = ‘auto’
• Max depth of trees = 8
• Split criterion = ‘gini’

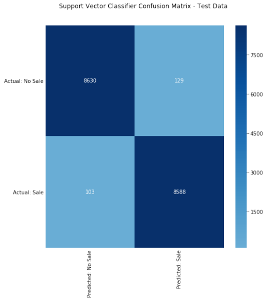
Support Vector Classifier
• C = range 0.06658900914614486
• Gamma = 0.22608939940417544
• Kernel = ‘rbf’

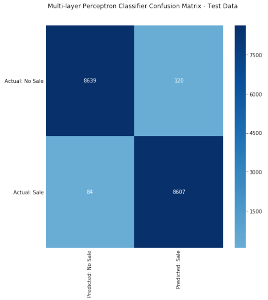
Multi-layer Perceptron
• Number of hidden layers = 3
• Number of hidden neurons = 100
• Activation function = ‘logistic’
• Learning optimizer = ‘lbfgs’
• Learning rate = ‘adaptive’

Performance Metrics

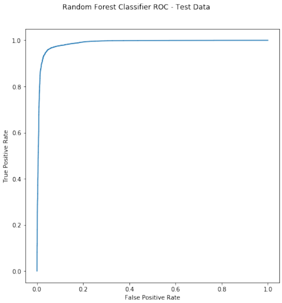
	Random Forest Classifier	Support Vector Classifier	Multi-layer Perceptron
Accuracy	95.3582%	98.6705%	**98.8309%**
Precision	0.9433	0.9852	**0.9862**
Recall	0.9648	0.9881	**0.9903**
F1-Score	0.9539	0.9867	**0.9883**
AUC	0.9874	0.9954	**0.9959**



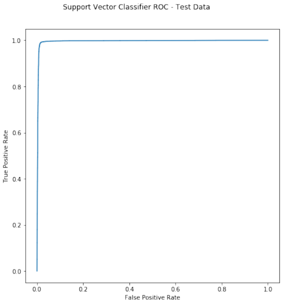




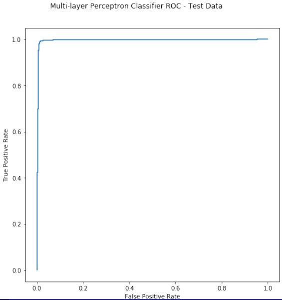
ROC Plots



Random Forest Classifier ROC - Test Data
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0



Support Vector Classifier ROC - Test Data
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0



Multi-layer Perceptron Classifier ROC - Test Data
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Insights

Optimal Machine Learning Model

- From the previous section, it is evident that the multi-layer perceptron is the best algorithm of the three models in terms of the performance metrics used for evaluation.
- Therefore, it can be concluded that the multi-layer perceptron should be the model of choice for predicting whether or not a customer will purchase a motor policy at Hastings Direct.
- Further improvements in performance could be achieved with extended hyper-parameter tuning (i.e. explore a wider set of hyper-parameters).

Key Findings

- The histograms in the outline and exploration section depicted the relationship between a customers age, the price they are willing to pay for a motor policy, and the frequency of customers that have and have converted a quote to a policy.
- From these plots, it was observed that customers aged between 40 and 60 are most unlikely to purchase a motor policy from Hastings Direct; this age group might not be willing to an average of £400 on motor insurance.
- On the other hand, customers aged between 18 and 40 are most likely to purchase a motor policy from Hastings Direct, paying £500 on average.
- Therefore, increasing marketing budgets, for example, on targeting customers aged between 18 and 40, and decreasing for customers aged between 40 and 60 could maximize income and yield the highest sales.

Change in Objectives

- The strategy mentioned on the previous slide would be optimal or near-optimal if Hastings Direct only objective was to maximize income and sales.
- However, Insurance is an industry of providing policies and also servicing claims from existing policies.
- By targeting a younger motorists (i.e. in the range of 18 to 40 years old) could potentially lead to an increase in claims.
- This is because young motorists are more likely to be involved in accidents than older, more experienced motorists (based on statistical evidence).
- Therefore, if Hastings Direct's main object was to minimize claims, and hence losses, targeting older motorists would most likely be successfully.

Thank you for listening!