Hastings Direct – Data Science Presentation Notes

Objective Slide

The process of predicting and/or projecting customer behaviour has always been a significant area of interest in the personal lines insurance market; through gaining intelligent insights into customer behaviour, insurers can tailor to customer needs and effectively meet their objectives. For example, the identification of loyal customers (i.e. customers that are most likely to provide repeat business) can help insurers price their insurance premiums appropriately.

The objective of this presentation is to identify a suitable machine learning model that can help predict whether or not a customer is likely to purchase a motor policy from quote data here at Hastings Direct. The machine learning models to be discussed in this presentation include the Random Forest, Support Vector Machine (SVM), and Multi-layer Perceptron (MLP).

Dataset Description

As previously mentioned, the dataset used to predict whether or not a customer will purchase a motor policy is based on historical quote data. The dataset consists of 50,000 observations with 11 predictor columns of which 7 are numerical, 3 are categorical, and 1 is a date. The values to be predicted are stored in the class column called `Sale'.

From initial observation, we observed that the dataset had a class imbalance; 30% of observations belonged to customers not purchasing a motor policy, and 70% belonged to customers purchasing a motor policy. The class imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE) algorithm. SMOTE, as the name suggests, is a over-sampling technique; however, instead of creating copies of observations in the minority class, it creates new observations in the neighbourhood of the existing minority data observations.

Furthermore, the dataset contained 5,496 missing values; there are two main methods for dealing with missing values. These include removing data observations containing missing values and using imputation methods to estimate missing values. For this specific task, we used an imputation technique known as Multiple Imputations by Chained Equations (MICE). MICE creates various copies of the dataset, uses various statistical methods to estimate missing, and pools the results together to produce final estimates.

Data Wrangling and Feature Engineering Slide

As machine learning models can only interpret numerical features, the 'Date' predictor was engineered into a meaningful numerical feature. The month of each policy inception date was extracted and stored into a new predictor column called 'Inception_Month'. The grouped bar plot shows the frequency of policies that were and were not purchased by month; this visualization shows that seasonality does have a possible relationship with a customers of purchasing or not purchasing a motor policy (i.e. customers seem to purchase more policies in early summer months). The categorical predictors (e.g. 'Marital_Status') were also label encoded to give them a numerical representation.

The grey matrix visualization depicts the positioning of the missing values that were contained in the dataset; as mentioned previously, these were addressed via the MICE algorithm.

Exploratory Data Analysis Slide

A correlation heat map was created to identify any correlations between the predictor columns and class column. As expected, the 'Price' predictor was positively correlated with the class column. The price of a motor policy can be considered as a main factor when deciding when or when to purchase a policy. The 'Age' predictor column was negatively correlated with the class column of the dataset; this indicates that young motorists are more likely to purchase motor policies from Hastings Direct in comparison to older motorists.

Boxplots of continuous-valued predictor columns were also plotted to identify any extreme outliers within the data. As depicted by the 'Credit_Score' boxplot, it was identified that this predictor contained several extreme outliers. A credit score is usually between 0 and 1000; the values above 1000 were set to null and re-estimated with the use of the MICE algorithm.

Histogram plots of the price distribution by age and class were also created to get a good idea of which age groups are more likely to purchase or not purchase motor policies. These results will be discussed further shortly.

Random Forest Description Slide

As previously mentioned, The Random Forest is one of the ML algorithms that has been implemented to solve our objective.

To understand how the Random Forest Algorithm works, we must understand the basic workings of a decision tree. A decision tree consists of a root node, terminal nodes, and leaf nodes. Using feature importance, a question is asked on the data of that feature. The decision made by the tree can lead to a leaf node (i.e. a prediction of class label) or terminal node (i.e. another decision is made on another feature that is contained in the dataset). Feature selection and placement is determined through a criterion, for example, the gini index.

The Random forest algorithm is constructed through an ensemble of decision trees based on a bootstrap aggregation of the training data. Bootstrap aggregation is simply the procedure of sampling random data observations in a training dataset with replacement. A subset of random predictors are selected for each tree to reduce correlation between each tree in the forest. Once the decision trees have predicted a class label, the class label that has the majority is used as the predicted class label.

Support Vector Machine Description Slide

Another ML algorithm used to solve our objective, is the Support Vector Machine (or Support Vector Classifier for classification tasks).

A SVM projects the data observations of a dataset into N-dimensional space; note that N usually is given by the number of predictors in the dataset. Once this has been done, the data is observed to see if it linear separable or not. If the data is deemed to be linear separable, a hyperplane is constructed to divide data observations from each class. The data observations closest to the hyperplane construct boundaries known as support vectors. The aim of the model is to identify a hyperplane that maximizes the distance between itself and the support vectors. If the data

observations are not deemed to be linear separable, a kernel function is used as the hyperplane (e.g. polynomial function).

Multi-layer Perceptron Description Slide

The last ML algorithm explored to solve our objective, is the Multi-layer Perceptron.

The MLP is a special type of artificial neural network that consists of a input layer, one or more hidden layers, and a output layer. Each layer consists of neurons which are formed of a summation function and a nonlinear activation function (e.g. sigmoid function). Each neuron in each layer is connected via a synaptic weight. The visual architecture of a MLP is similar to that of a lattice.

The MLP algorithm is usually trained through an optimization method known as backpropagation. Back-propagation consists of two phases: the forward phase and the backward phase. The forward phase consists of producing meaningful predictions using the neurons in each layer. The prediction and the actual class label are compared via a loss function (e.g. binary cross entropy for a binary classification task), and the error is propagated back through the network and the synaptic weight values are adjusted accordingly using stochastic gradient descent. This process is repeated until the loss function is minimized or a certain number of passes through the network have been completed.

Training and Evaluation Methodology Slide

For this experiment, the dataset was split into training and test set using a 75/25 % split. A random search was conducted on each model to source 'ideal' hyper-parameter values for that model. 5-fold cross validation was also used on the training set whilst training the models to mitigate the risk of overfitting.

The evaluation metrics used to assess the models' performance included the accuracy, precision, recall, and f1-score. These metrics were calculated from true and false values contained in a confusion matrix. A roc curve was also plotted to identify which model had the best area under the curve score.

The hyper-parameters considered for the Random Forest algorithm included the number of trees in the forest, the number of features to consider when splitting a root/terminal node, the max depth of the trees in the forest, and the split criterion.

The hyper-parameters considered for the SVM algorithm included the value of the penalty term of misclassifying observations, known as C, gamma, and the kernel of the hyperplane separating data observations in each class.

The hyper-parameters considered for the MLP included the number of hidden layers in the network, the number of neurons in each hidden layer, the activation function to be used in each neuron, the best optimization algorithm available to conduct backpropagation in the sklearn package, and the learning rate (the method to determine the amount the synaptic weights should be adjusted).

Best Model Hyper-Parameters Slide

The best hyper-parameters that were sourced during the random search are shown.

Performance Metrics Slide

The performance metrics that were previously mentioned are shown for each model. We observe that the Multi-layer Perceptron was the best ML model at classifying whether or not a customer would purchase a motor policy from Hastings Direct. The Support Vector Machine was the second-best algorithm, and the Random Forest algorithm was the least effective (although still had exceptional performance).

The confusion matrix for each model is shown at the bottom of the slide. Furthermore, the ROC curves are shown on the next slide. It is evident that the MLP's true positive rate converges to the value 1 the quickest, thus giving it the best AUC score.

Optimal Machine Learning Model Slide

From the performance metrics, it is advisable that Hastings Direct could potentially train and deploy a Multi-layer Perceptron to classify customers in respect to whether or not they are likely to convert their motor quote to a policy. As the MLP in this experiment was constructed with limited time, further improvements in model performance could potentially be achieved via extended hyper-parameter tuning.

Key Findings Slide

Going back to the price distribution histograms depicted in the previous EDA slides, it was observed that customers aged between 40 and 60 are most unlikely to purchase a motor policy from Hastings Direct. The price of the policies could be the main contributors to this result; customers aged between 40 and 60 might be unwilling to pay roughly £400 on average for motor insurance.

On the other hand, customers aged between 18 and 40 are most likely to purchase motor policies from Hasting Directs. Customers aged between 18 and 40 are willing to pay roughly £500 on average for motor policies.

Therefore, increasing advertising and marketing budgets, for example, for customers aged between 18 and 40, and reducing for customers aged 40 and 60 could potentially maximize income and sales.

Change in Objectives Slide

The strategy of targeting customers aged between 18 and 40 would only be good if Hastings Direct only objective was to maximize income and sales. However, increasing exposure to younger motorists could lead to a substantial increase in claims as young drivers are more likely to be involved in accidents (based on statistical evidence). If Hastings Direct had the objective of minimizing claims, targeting older drivers could be a feasible strategy as they have more motor-related experience on average.