Digital Marketing Analytics Coursework 2 – Question 2

Pre-processing (to both train set and test set)

1.Drop the rows that contain NA;
2.Drop the rows whose 'firstorder' column equals obviously wrong values such as '1/0/00', '1904-01-01' and Timestamps that equals 0;
3.Convert columns 'firstorder', 'lastorder' and 'created' into Timestamp objects;
4.Convert columns 'city' and 'favday' into dummy variables

Variable creation

1.Variable 'order_duration' is created using the difference between the time of first order and of last order
2.Variable 'recency' is created using the quantile of each 'lastorder' among all 'lastorder' data. The whole is segmented into 5 equal-size levels. The closer the time of last order is to now, the higher the 'recency'.

Models we have tested

	LASSO	Elastic Net	Random Forest	K-Nearest Neighbours
Training Accuracy	0.9341	0.8988	0.9586	0.9385
Test Accuracy	0.9383	0.8974	0.9580	0.9325

Random forest is the best – and why

1.Compared to many other machine learning methods, random forest is way more interpretable as we can extract features with their order of importance in the random forest;
2.This customer churn case is a classification problem so random forest stands out due to its nature of classification;
3.There exists high multicollinearity problem within our dataset. The multicollinearity problem can be alleviated by random forest since a random subset of features is chosen for each tree which is an advantage over simple tree models;
4.The data available is not temporal or sequential data - each row is the information of one customer. Therefore, the popular neural network does not have much advantages here (especially considering the computing time problem).

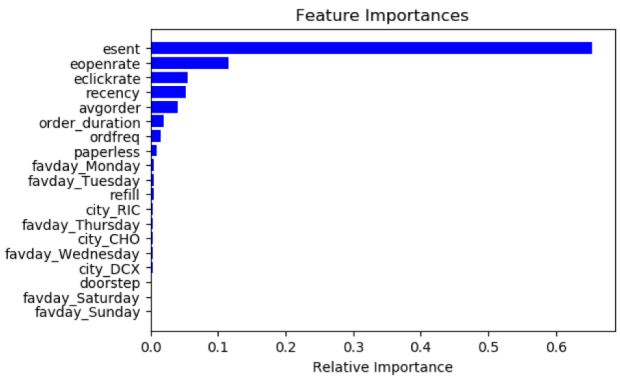
Hyperparameter tuning & tradeoff between overall accuracy and false-negative rate

Based on the feature engineering and the model selection, we executed the grid search on the Random Forest, which is a method to find a model with the best performance by tuning the hyper-parameters in the selected model. As a result of tuning, the confusion matrix shows 94.3% of the model accuracy, which demonstrates the reasonably high performance. However, in view of the churn management in business, it is more important to minimise the false negative, which indicates that the classier failed to identify the customers who will churn, as the loss of losing a customer is generally much higher than the cost of taking actions to save back a customer (even though he/she might not churn at all). When the threshold to the predicting probability is 0.64, the model performance is even better in the false negative rate and the accuracy as below.

Threshold : 0.5 (Default) Accuracy : 94.3%		Predicted	
		Churned	Retained
Actual	Churned	940	341
	Retained	9	4918

Threshold : 0.64 Accuracy : 94.8%		Predicted	
		Churned	Retained
Actual	Churned	984	297
	Retained	25	4902

Variable importance



The Random forest model can show the variable importance which is measured by the decrease of the misclassification when split by each predictor. As depicted on the left, the plot indicates that the Number of emails sent ("esent") is by far the most critical variable regarding the model accuracy. In other words, to reduce the churn rate, the client marketing team can increase a budget for frequent emailing to customers.