Project Jarvis

I. INTRODUCTION

In today's world, online payments, and online payment systems are becoming more ubiquitous. Consequently, as more items and services are bought and paid for with credit and debit card services, fraudulent transactions that relate to it, similarly, are on the rise. Fraud, here, is defined as deception intended to result in personal gain. For online business, this often takes the form of compromised credit cards, debit cards, or user accounts. Since fraud itself is not the common case [1], the modeling and detection of fraud has become a very important area of research.

The data was provided by The Hut Group, an e-commerce company that manages over 100 websites and sells worldwide. All their business is done online, and there is a risk that any order could be fraudulent. They use an automated system that refers suspicious orders to investigators, who decide which of those orders to reject. While it is important that true cases of fraud are sent to the investigators, the system should also approve the vast majority of non-fraud cases automatically, so that business is not bogged down by a large queue of orders to be checked.

The aim of this project is to establish a key set of variables that well categorize the problem of fraud. Also, to investigate whether a general model aptly classifies transactions, or whether it would have to be created to be site, region or locale specific.

To answer the above questions, through extensive data exploration, a number of features were engineered from the original data set in a way that more aptly categorizes the occurrence of fraud. Following feature selection and engineering, the data which were provided by The Hut Group company, were also standardized, normalized and categorized as necessary, in order to prepare the variables as input for logistic regression models, and for classification models. The classification models employed include a Naïve Bayes classifier, and the ensemble Random Forest classifier. For the logistic models created, feature significance was established through methodical backward elimination based on the Akaike Information Criterion (AIC), which penalizes model complexity, as well as provides a means by which models' effectiveness can be compared [2].

Following this, model validation was done by means of ten-fold cross validation, and analysis of the area under the models' respective Receiver Operating Characteristic (ROC) curves. Finally, the models, and their respective defining features were examined holistically and apparent inferences were drawn from the overall procedure of fraud detection; the results revealed that the key features which were created from the original data set are significant across all models.

II. METHODOLOGY

A. Research Strategy

The first step of the research strategy was to create a general model by which a transaction may be labeled as fraudulent or otherwise. The data set was split based on site, region and locale.

It was then examined to see whether the general model could make accurate predictions for the split data or if it was necessary to build new models based on a new set of variables. Since the outcome variable was binary, logistic regression model was an apt model to apply to this problem. The Naïve Bayes and Random Forest classifiers were also used because they also map well to the domain of classification problems around financial fraud involving categorical or binary data [3]. The above models were then compared.

B. Preprocessing

The main data consisted of three CSV files containing transactions, customer details and Chargeback details. Seven additional CSV files were

also provided and contained the encoding of the following categorical variables *Delivery Option Type Key*, *Country Code*, *Payment Provider Key*, *Payment Method Key*, *Locale Key*, *Medium Key* and *Payment Method*. It was found that the values of some features were in the wrong column, or split across multiple columns. This was corrected manually, and the empty columns *X*, *Empty 1* and *Empty 2* were removed. The data sets were then merged and formatted using R, by their foreign key variables *Order Number* and *Account Key*. The data set originally contained over fifty features. In order to reduce its dimensionality and extract pertinent information, a new data set was created, with newly engineered features.

The Chargeback data contained a variable, *Internal.RC*, which described the conclusion of an investigation into the Chargeback. A fraudulent transaction was therefore defined as one whose *Internal.RC* was indicative of fraud. The binary variable *isFraud* was constructed from this criterion; 1 for fraud, and 0 for a non-fraudulent activity.

The variable *Customer IP Address* was transformed into a new one, *IP Country*, which contained the countries corresponding to the IP addresses the transactions originated from.

The variables *Registered Date* and *Order Date Key* were replaced by the variable *Account Age* which represents the age of the account in months, based on the time that had passed from the registration date till the date of the transaction. Similarly, the variable *Active Age* was constructed to represent the time (in months) that had passed from the date of the first order (*First Order Placed*) till the date of each transaction (*Order Date Key*). The following eight binary variables *reg ship same*, *reg bill same*, *ip reg same*, *ip bill same*, *ip ship same*, *isPcSame1*, *isPcSame2*, *isPcSame3*, *isCcSame* were also created in order to check if the shipping, billing, registration and IP address origin countries and postcodes were identical; 1 for true, 0 for false.

As shown in Figure 1, among the product categories represented by the variable *Category Level 2*, games, clothing and footwear had the highest frequency of fraudulent transactions. So, *Category Level 2* was transformed into a new categorical

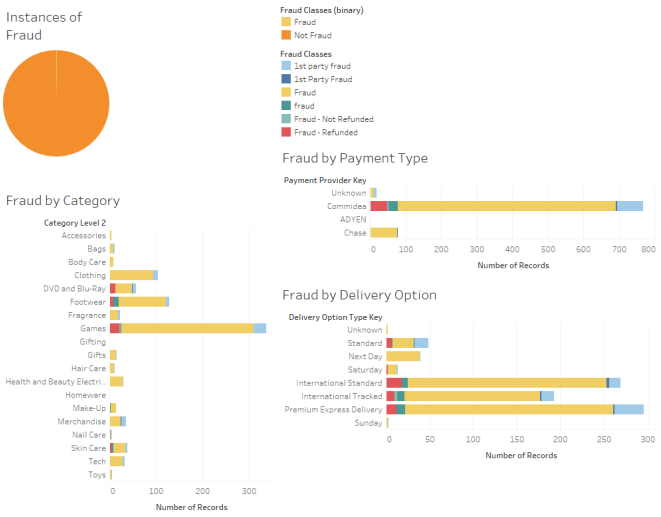

Fig. 1. Fraud Distribution

variable with the three levels for games, clothing/footwear and others. It was also observed that fraudulent transactions did not contain any Paypal payment (Figure 1), necessitating a new binary variable, *isPaypal*, which represents whether the payment method was Paypal or not.

The variable *Medium Key* was replaced by the following two binary ones. It was observed that customers who came to the site by entering explicitly the url address (directly) or by searching for it using its name (organically) tended to commit more fraud, which was represented by the newly engineered binary variable *isDirectOrganic*. On the other hand, those who accessed the site through their email or a paid personal site (affiliate) were less likely to commit fraud. This information was represented by the binary variable *isMailAffiliate*. Fraudulent transactions were also discovered to be more probable for products which are on campaign. The variable *Campaign Key* was replaced by the binary one *isCampaign* which represents whether some or all products within the order were sold as part of a campaign or not. In Figure 1, the delivery types *International Standard*, *International Tracked* and *Premium Express Delivery* were indicative of potential fraud. Therefore, the binary variable *delivery* was constructed to represent whether the *Delivery Option Type Key* belonged to one of the above categories or not. The continuous variable *Product Charge Price* which represents the price of the product affected by any

discounts was kept in the data set. The variables *Locale x*, *Site Key*, *Country Name Shipping* and *Continent Shipping* were maintained in the data set for later use; they formed the basis of the data split by locale, site and region.

Since exploratory analysis showed that the number of fraudulent transactions for the English speaking locales was remarkably high in comparison with the rest, the locales *en GB* and *en US* were examined. All the three sites included in the data set were examined and had the following codes *11*, *15* and *121*, *120*, *153*, *119* respectively. The regions examined were the Americas, which includes countries from North, Central and South America, EMEA, which includes countries from Asia, Eurasia, Europe and Africa, and the United Kingdom. The latter constituted a unique category because it was observed that the instances of fraud in this country are significantly great. The exploratory data analysis for the rest of the variables included in the data set, showed little relation to the response variable *isFraud*, so they were removed from the data set. In addition to this, it was found that there were seven missing values, that is 0.000017 of the data. Since this is a very small percentage, these transactions were removed from the data set. In order to handle the outliers, the continuous variables *Product Charge Price*, *Account Age* and *Active Age* were standardized. This procedure resulted in 12,840 out of 418,396 transactions lying out of the normal range. The ensuing analysis on the outliers indicated some consistency with the rest of the data, so they were maintained, in order to keep as much information as possible.

C. Analysis

Some descriptive statistics were also sought in order to gain insight for the nature of the variables. The data set included 866 fraudulent transactions and 417,530 non-fraudulent ones. This results in a proportion of the non-fraudulent transactions to the fraudulent of approximately 482:1, indicating a very clear imbalance [4]. Specifically, the minority class is the fraud and the majority class is the no fraud. In order to handle the class imbalance problem, undersampling of the majority class was applied to enforce some data balance [5], [6]. Finally, the 866 fraudulent transactions were kept in the data and 866 non-fraudulent ones were randomly selected. The logistic regression model was initially used to find an appropriate model to fit the data. Backward elimination based on the Akaike information criterion (AIC) was applied for the variable selection [7]. Logistic regression estimates the log-odds of p_i as

$$log\left(\frac{p_i}{1-p_i}\right) = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_\rho x_{\rho i}$$

where p_i is the probability that observation i is fraud, $x_{1i}, \ldots, x_{pi}$ are the values of the variables and $\beta_0, \ldots, \beta_p$ are the estimated coefficients [8]. Nagelkerke pseudo-R^2 was used to assess the model. The Naïve Bayes and random forest classifiers were also performed after normalizing the continuous variables using the formula

$$Z = \frac{X - min(x)}{range(X)}$$

The naïve bayes model estimates the probability $P(\text{Fraud}|x_{1i}, \ldots, x_{pi})$ with Bayes' theorem, making the simplifying assumption that the covariates are independent of each other, so that the algebra simplifies to

$$\frac{P(\text{Fraud}|x_{1i}, \ldots, x_{pi})}{P(\text{Not Fraud}|x_{1i}, \ldots, x_{pi})} = \frac{P(\text{Fraud})P(x_{1i}|Fraud) \cdots P(x_{pi}|Fraud)}{P(\text{Not Fraud})P(x_{1i}|Not\ Fraud) \cdots P(x_{pi}|Not\ Fraud)}$$

$P(\text{Fraud})$ can be estimated as the overall proportion of cases that are fraud [9].

A random forest is a large collection of decision trees. Each of these trees is grown from a randomly selected subset of the variables. For observation i, p_i is estimated as the proportion of trees in the forest that predict the observation to be fraud [9].

The undersampling data were divided into two parts. 75% of the data was used to train both the logistic regression model and the two classifiers, as well as to overcome overfitting problems, with the repeated 10-fold cross validation as Geoffrey et al. suggest [10]. The data were randomly shuffled five times and each time they were split into training and testing data using 10-fold cross-validation. The remaining 25% was used as new data to test the predictive power of the model, which was also done by site, locale and region. The three models

were compared by their accuracy, confusion matrix, along with the corresponding ROC curve and the index Area Under Curve (AUC).

III. RESULTS

The stepwise backward elimination based on the AIC for the logistic regression showed that the most informative variables concerning the fraudulent transactions to be the following: *Account Age*, *reg ship same* , *ip ship same* , *Active Age* , *isPcSame1*, *isPcSame2*, *isPcSame3*, *delivery*, *Category Level 2*, *isPaypal*, *isDirectOrganic* and *isMailAffiliate*. The AIC value for the final model was 1181.1, while its initial value was 1188.4 and the Residual Deviance was equal to 1153.1, with 1696 degrees of freedom. It was found that the model interprets 74% of the overall variance (Nagelkerke pseudo-R^2 = 0.74). It was calculated that the Pearson's correlation between *Account Age* and *Active Age* is 0.96. After experimentation, it was found that eliminating *Account Age* from the model increased the accuracy, so it was dropped in favour of *Active Age*. The phi correlation coefficient between *isPcSame2* and *isPcSame3* was equal to 0.86 and after experimentation similar to the above, *isPcSame3* was also dropped from the model.

The final resulting general model was of the form:

$$log\left(\frac{p_i}{1-p_i}\right) = \eta_i = \beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+...+\beta_\rho x_{\rho i}$$

or,

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+...+\beta_\rho x_{\rho i}}$$

$$\frac{p_i}{1-p_i} = e^{\beta_0}\cdot e^{\beta_1 x_{1i}}\cdot e^{\beta_2 x_{2i}}\cdot ...\cdot e^{\beta_\rho x_{\rho i}}$$

from which the probability p_i is determined as,

$$p_i = \frac{e^{\beta_0}\cdot e^{\beta_1 x_{1i}}\cdot e^{\beta_2 x_{2i}}\cdot ...\cdot e^{\beta_\rho x_{\rho i}}}{1+e^{\beta_0}\cdot e^{\beta_1 x_{1i}}\cdot e^{\beta_2 x_{2i}}\cdot ...\cdot e^{\beta_\rho x_{\rho i}}}$$

The coefficients estimates for the above model are presented in the Table I. It is worth noting that the coefficient for *isPaypal* has the greatest absolute value of all the considered variables. This means that it has the strongest individual effect in the response variable. It was calculated that there is an approximate 99.99% relative reduction in odds for the customers who pay with Paypal versus the odds for those who do not. In contrast, the coefficient of *Active Age* had the smallest absolute value, and hence, the least individual effect on the response variable. It was calculated that a one-unit increase in *Active Age* results in an approximate 6.4% reduction in the relevant odds.

TABLE I
MODEL COEFFICIENTS

Coefficients	Estimate
Intercept	-0.89878
reg ship same1	2.26397
ip ship same1	-0.51249
Active Age	-0.06612
isPcSame11	-0.96212
isPcSame21	0.62153
delivery1	1.72459
Category Level 22	-0.32966
Category Level 23	-1.63461
isPaypal1	-18.27755
isDirectOrganic1	1.29500
isMailAffiliate1	-0.83777

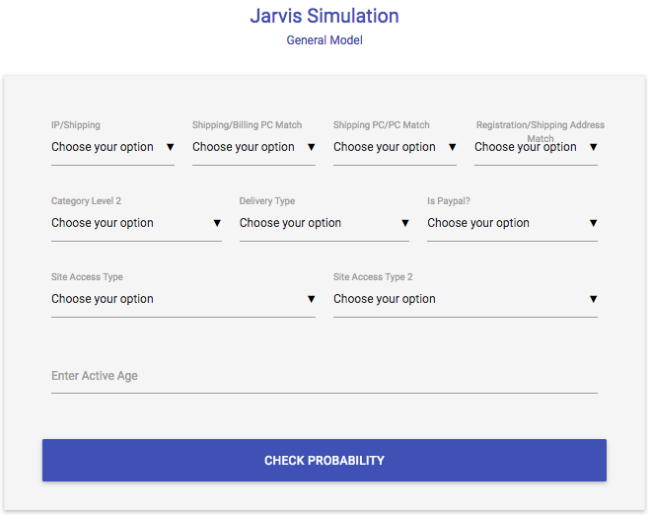

Fig. 2. regression simulation

Figure 2 shows a logistic regression simulation currently hosted at http://www.pkopoku.github.io.

The Logistic Regression, Nave Bayes and Random Forest models were trained and tested with repeated 10-fold cross-validation.

As shown in Table II, the variables selected are very informative concerning the instances of fraud.

TABLE II

GENERAL MODEL RESULTS

	Logistic Regression			Naïve Bayes			Random Forest		
General model	Train Accuracy	Prediction Accuracy	AUC	Train Accuracy	Prediction Accuracy	AUC	Train Accuracy	Prediction Accuracy	AUC
	85.9%	86.52%	**94%**	86.22%	89.36%	**94.5%**	89.65%	90.78%	**93.9%**

	Logistic Regression		Naïve Bayes		Random Forest	
	Training Set	Testing Set	Training Set	Testing Set	Training Set	Testing Set
Sensitivity	85.9%	90.3%	83%	86.9%	85.2%	85.9%
Specificity	**85.8%**	**82.8%**	**89.2%**	**91.6%**	**93.9%**	**95.3%**

Consequently, the accuracy of the general model is high for all the models. Moreover, it can be noticed that there is a minor difference between the accuracy of the training and the testing data set. However, this can be easily ignored as the rate is very low. One reason behind this, can be that the training set, had many fraudulent transactions to learn, so it was easier for the testing set to predict the fraudulent behavior. The random forest was the most appropriate classifier for this problem since its accuracy value of 90.78% is the highest.

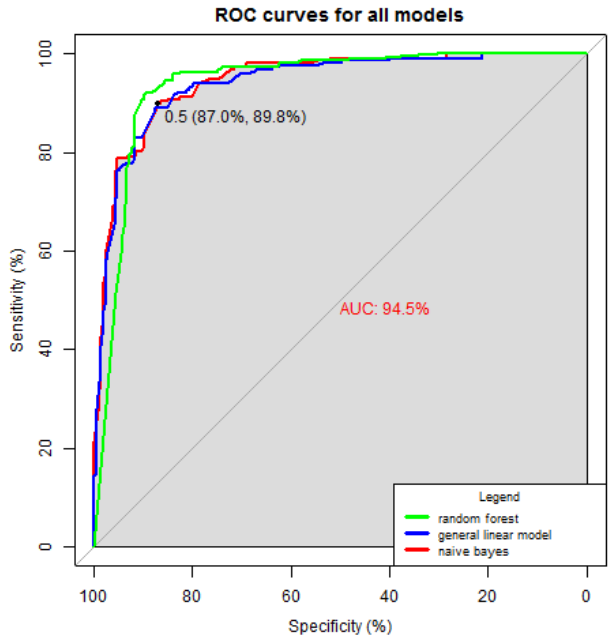

Fig. 3. ROC Curves

The specificity percentages by definition depict the true negative values. Consequently, in Table II, they describe the rate of the correct labeled features, as fraud. These rates are fluctuating from approximately 83% to 95%. The percentages, are very high and reveal low misclassification errors. Furthermore, it can be concluded from the Figure 3, that the AUC rates are notably high, indicating that all the three models can adequately solve the problem. Specifically, the highest value 94.5% derives from the Nave Bayes classifier.

Subsequently, it was investigated if the set of the variables that the regression model indicated, is sufficient to identify the fraudulent transactions by site, region or locale. To this end, 25% of the undersampled data was filtered by site, region and locale in order to test the predictive power of the models.

A. Results per Site

TABLE III

SITE RESULTS

	Logistic Regression			Naïve Bayes			Random Forest		
Sites	Prediction Accuracy	AUC	Specificity	Prediction Accuracy	AUC	Specificity	Prediction Accuracy	AUC	Specificity
11 & 15	76.4%	76.4%	75.3%	88.4%	88.4%	90.2%	90.5%	90.7%	92.1%
153-119-120-121	89.7%	95.3%	86.7%	90%	95.6%	91.6%	91.2%	94.7%	91.6%

Due to the undersampling, the testing data for Site 11 and 15 ended up to be 37 and 65 observations respectively. For this reason, it was decided to aggregate these two groups and make predictions for both sites, using in total, 102 observations for the testing set. The results of the prediction are presented in Table III.

The accuracy of the predictions fluctuates incrementally between 77% and 92%. The lowest accuracy level belongs to the logistic regression, whereas the highest to the random forest. In addition to the above, the specificity levels are high and approximately similar to the values corresponding to the general model. Concerning the AUC percentages, all of them are also high; specifically between 76.4% and 95.6%, which means that the chosen models address the fraud detection problem sufficiently.

B. Results per Region

As shown in Table IV, the accuracy of the predictions for all regions and particularly for the UK is increased in comparison to the values depicted in Table II. The lowest accuracy level belongs to the logistic regression and specifically pertains to the region America, whereas the highest is matched to the random forest. Furthermore, specificity as well as AUC levels are high, which indicates that

TABLE IV

REGION RESULTS

	Logistic Regression			Naïve Bayes			Random Forest		
Regions	Prediction Accuracy	AUC	Specificity	Prediction Accuracy	AUC	Specificity	Prediction Accuracy	AUC	Specificity
UK	93.3%	95.7%	88.2%	92.3%	96.9%	91.4%	92.8%	97.1%	89.3%
AMERICA	78.6%	80.7%	81.6%	87.6%	81.3%	92.9%	88%	80.5%	93.5%
EMEA	83.5%	93.3%	79.1%	85.5%	93.7%	89.5%	88.9%	92%	93.7%

the selected models are able to identify fraudulent behavior for each region.

C. Results per Locale

TABLE V

LOCALE RESULTS

	Logistic Regression			Naïve Bayes			Random Forest		
Locales	Prediction Accuracy	AUC	Specificity	Prediction Accuracy	AUC	Specificity	Prediction Accuracy	AUC	Specificity
en_GB	86.2%	94.1%	81%	89.5%	94.9%	89.1%	89.5%	94.9%	89.1%
en_US	87.5%	93.3%	93.3%	93.7%	93.3%	100%	93.7%	93.3%	100%

As shown in Table V, the accuracy for both locales and for every classifier is increased. The same pattern follow the Area Under the Curve's results. These rates were expected, since there is a resemblance between fraudulent transactions in English speaking locales. However, it can be noticed that the specificity levels for *en US* locale using Naïve Bayes and Random Forest classifier is 100%. This is very likely the case because the data used for testing the predictive power of the models include very few observations for this locale.

D. Inferences

All things considered, it can be summed up that the global set of variables that was used for the general model, is sufficient for the detection of fraudulent transactions. This global set, after being tested on different subsets of data based on the site, region and locale, had high accuracies across the data splits, and had little differences compared to the general model. Consequently, it was not necessary to build new models for each site, locale and region, as a result of the features that were engineered for the models in the preprocessing stage. All the variables that were used in the final general model, are a combination of the original variables in the initial data set. Finally, the only trend for fraud that was identified, is in the UK region. This result was expected, because based on the fraudulent transactions that were analyzed in depth, the biggest proportion of them were coming only from this country.

E. Bias and Validity

It is worth mentioning that the modeling construction was based on a specific set of variables provided by The Hut Group, which assumes that the data provided is sufficient for fraud detection. However, another important variable, which was not included in the data set, is the spending behavior of the customer, which is anticipated to contribute to the investigation of fraud patterns [11]. A widely recognized and used way so as to take into consideration the aforesaid variable, is by following a transaction aggregation strategy [12]. Moreover, since the data refer only to a specific three-month period (Spring), and only to three sites, the results are valid only for these particular sites and time period. Concerning external validity of the study, although the models which were constructed gave accurate predictions for the undersampled data, they did not perform well for the original data set.

IV. CONCLUSION

This research investigated the key set of variables required to identify fraudulent transactions, based on the data provided by The Hut Group. In addition, it determined whether these variables well capture fraudulent behaviour, or whether region, locale and/or site specific variables were needed to identify fraud. Following a thorough data exploration and preprocessing, logistic regression was used to construct the final model and based on its variables, Naïve Bayes and random forest classifiers were trained and tested. The outcome of this study indicates that the key features which were created from the original data set, are significant across all models, as the accuracy levels are considerably high, obviating the need to create extra models per site, region and locale. The only noteworthy difference in fraudulent transactions between different regions was for the UK, since the most incidents of fraud were coming from this country. These findings of the study are promising

and could pave the way for the building of an Artificial Fraud Investigator.

However, there were some emergent limitations encountered. It was found that the constructed model gave accurate predictions per site, locale and region using the 25% of the undersampled data. However, the accuracy of this model's predictions for the original data, excluding the data used for training decreased significantly. Given that the proportion of fraudulent and non-fraudulent transactions in the undersampled data was 1:1, more experimentation for the proportion of fraudulent and non-fraudulent can be done so that the results improve. Further still, alternative methodologies, such as oversampling of the minority class, both oversampling and undersampling, and bootstrapping, to overcome the class imbalance problem could be adopted [11]. The variables which were found to be informative for the logistic regression model using backward elimination based on AIC, were also used for the Naïve Bayes and random forest classifiers. Nevertheless, more extensive and thorough processing is required for better selection of the variable set in order to improve the results for both classifiers. Finally, Weight of Evidence could be used to code the predictor variables and information value could be applied for variable selection [5]. It would also be interesting if data were collected from another period of time, such as black Friday, when more transactions in general are fraud [13], [14], [15].

ACKNOWLEDGMENT

The Datanauts are deeply appreciated of the support extended by The Hut Group, and Yehia Elkhatib.

REFERENCES

[1] K. Chaudhary, J. Yadav and B. Mallick. *A review of Fraud Detection Techniques: Credit Card*. International Journal of Computer Applications, Volume 45 No.1, May 2012.

[2] E. Wagenmakers and S. Farrell. *AIC model selection using Akaike weights*. Psychonomic bulletin and review, 2004 - Springer

[3] 2. C. Liu , Y. Chan , S. Kazmi and H. Fu. *Financial Fraud Detection Model: Based on Random Forest*. International Journal of Economics and Finance; Vol. 7, No. 7; 2015.

[4] R. Lima and A. Pereira. *A Fraud Detection Model Based on Feature Selection and Undersampling Applied to Web Payment Systems*. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE (2015).

[5] Weis et al. *Cost-Sensitive Learning vs Sampling*.2007.

[6] P. Brennan. *A Comprehensive Survey of Methods for Overcoming the Class Imbalance Model in Fraud Detection*. Institute of Technology, Blanchardstown, Dubline, Ireland. June 2012.

[7] E. Shtatland, E. Cain, and M. Barton. *The Perils of Stepwise Logistic Regression*. Harvard Pilgrim Health Care, Harvard Medical School, Boston, MA. 2008.

[8] McCullagh, P. and Nelder, J.A. *Generalized Linear Models, Second Edition*.Chapman & Hall/CRC

[9] Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome. *The elements of statistical learning: data mining, inference and prediction*. Springer. 2009.

[10] G. McLachlan, K. Do and C. Ambroise. *Analyzing microarray gene expression data*. Wiley 2005.

[11] S. Bachmayer. *Artificial Immune Systems*. Vol.5132,pp.119131.doi:10.1007/11823940. 2007.

[12] Whitrow,C.,Hand,D.J.,Juszczak,P.,Weston,D.J., and Adams,N.M. *Transaction aggregation as a strategy for credit card fraud detection*. Data Mining and Knowledge Discovery. Springer. 2008.

[13] *Industry Perspective, Peak Trading - The Perfect Storm For Fraudsters*. ACI Universal Payments, 2016

[14] *Online Payment Fraud Whitepapaer*. Juniper Research, 2016.

[15] *Global Risk Technologies, Black Friday - Golden Payday or Poisoned Chalice?*. Nov 2015.