International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

Vol. 32 No. 2 May 2022, pp. 69-78

Modelling Credit Defaults Using Support Vector Machine And Binary Logistic Models

Peter Gachoki[1], Lucas Macharia[1] and [1]Kinyanjui Jeremiah Ndung'u

[1]Department of Pure and Applied Sciences, Kirinyaga University,

P.O Box 143-10300, Kerugoya, Kenya

Abstract – Defaulting on a loan essentially occurs when an individual has stopped making payments on a loan or credit card according to the account's terms. A default model is constructed by financial institutions to determine default probabilities on credit obligations by a corporation or sovereign entity. A probability of default model uses multivariate analysis and examines multiple characteristics or variables of the borrower, and it will usually account for credit or business cycles by either incorporating current financial data into the generation of the model or by including economic adjustments. Modelling loan default allows financial institutions to determine typical features and patterns of behavior that lead to a future inability to make debt repayments. This modelling helps to assess the probability of future default for each client. The focus of this study was to apply the support vector machine and binary logistic models to model credit defaults. The process involved identification of the predictors that could be associated with credit defaults as well as comparison of the performance of the prediction models on their statistical power to model credit defaults. The analysis was done using R statistical software. The results showed that variables; credit amount, marital status, credit history and location of property used as security were significant predictors of credit defaults. The results also showed that the binary logistic model had a better performance that the support vector machine model in terms of F1 score and accuracy of predicting credit defaults. The logistic model had the accuracy of 0.826087 and an F1 score of 0.8809524. The support vector machine had the accuracy of 0.7826087 and an F1 score of 0.8554913. From the study findings, it was concluded that, the accuracy of the prediction models in modelling of credit defaults was dependent on the variables considered. Different set of variables would yield different accuracies for the prediction models.

Keywords – Credit defaults, support vector machine, binary logistic, accuracy, F1 score.

I. INTRODUCTION

A default occurs if the lender decides to an account due to missed payments (Ganong & Noel, 2020). This might happen to an account one has with a bank, mobile phone company or utility supplier. A default can occur regardless of how much money is owed, whether it's a few pounds or a few thousands. It usually happens if payments have been missed over the course of three to six months, but this can vary depending on the lender's terms. Unpaid loan or interest balances has consequences such as lack of deferment or forbearance, lost eligibility for other benefits, such as the ability to choose a repayment plan. To the lender it can also lead to losses (Brooks & Levitin, 2020).

Mathematical modelling is way of making approximations from input data. These approximations are then used to make predictions (Lunt, 2013). The developed models help in predicting the future probabilistic behaviour of a system based on past statistical data (Geisser, 2016). Predictive modelling has been used in many fields, for example in crime cases (Finlay, 2014); to detect the likeliness of an email being spam (Sheskin, 2011) and in modelling credit defaults. Credit deafult modeling is the application of statistical models to creditor practices to help create strategies that maximize interest and minimize defaults. Credit default models are used to quantify the probability of default or prepayment on a loan (Glennon & Nigro, 2011).

Corresponding Author: Peter Gachoki

Several variables have been considered by researchers when modelling credit defaults (Bach *et al*., 2017). For instance, on analysis of the financial health of a company taking credit, it was found that companies with large income had low tendency of defaulting contrary to those which have small income whose probability of default is high. On age of the company account, it was found that there's a decreasing default trend in conjunction with account longevity except for those companies whose account ages are less than 3 years. Other variable that have been used in modelling individual credit defaults include; monthly expenditure in most recent 12 months, number of dependents, age, additional income, primary income, employment status, average expenditure and duration employment. In this study, the variables considered were the; loan amount, marital status of the borrower, credit history of the borrower and location of the property used as security (Bach *et al*., 2017).

Several models have also been used to model credit defaults. For instance, survival analysis has been used to study influence the company's repayment performance. The study findings showed that oldest companies whose accounts were opened more than 8 years before loan application had a lower tendency of default (Masai, 2020). The study also showed that Nelson Aalen was a better estimator of time to default to Kaplan-Meier. The GEV regression model has been used to model loan defaults in Kenya banks. The results of GEV were compared with the results of the logistic regression model. The study found out for rare events such as loan defaults the GEV performed better than the logistic regression model (Wanjohi et al., 2016). As the percentage of defaulters in a sample became smaller the GEV model to identify defaults improved whereas the logistic regression model becomes poorer. In the use of statistical and machine-learning classifiers to explore the suitability of cooperative models and bootstrapping strategies for default prediction, the results indicated that combinational approaches based on correlation-adjusted strategies are promising techniques for managing sparse LDPs and providing accurate and well-calibrated credit risk estimates. The modelling of credit defaults is motivated by availability of data (Florez & Ramon, 2014).

Data has the potential to transform business and drive the creation of business value. It can be used for a range of tasks such visualization relationships between variables to predicting if an event will occur (Edwards, 2019). The latter is one of the heavily researched areas in recent times. The reason for this is that data has grown exponentially and so has the computing power. Banks and financial institutions have used data analytics for a range of value such as fraud detection customer segment, recruiting, credit scoring and so on. In this study a data set was used to build a credit model. The main motivation was to build a model that minimized as much as possible the number of false positives and the number of false negatives. This is because predicting false positives will eventually cause a business to make a loss and false negatives means that the financial institution loses business. The models for this study are the support vector machine model and the binary logistic model.

II. METHODOLOGY

2.1 Data preprocessing and Variables Visualization

The data preprocessing involved data cleaning and encoding of the categorical variables. Data cleaning involved the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within the dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled and this warranted the data cleaning. Data coding of data involved the process of transforming collected data to a set of meaningful, cohesive categories. It involved summarizing and re-presenting data in order to provide a systematic account of the recorded or observed phenomenon. In this study, the one hot encoding was applied to code the data. One hot encoding is a way of preprocessing categorical features for machine learning models. This type of encoding was used to create a new binary feature for each possible category and assign a value of 1 to the feature of each sample that corresponds to its original category. Data visualization involved the representation of data through use of common graphics, such as charts and plots. These visual displays of information helped to communicate complex data relationships and data-driven insights in a way that was easy to understand. The variable visualization in this study was done using stacked bar plots. A stacked bar plot was used to break down and compare parts of a whole. Each bar in the chart represented a whole, and segments in the bar represented different parts or categories of that whole. Box plots was used to visualize the numeric variables.

2.2 Binary Logistic Model

Binary logistic regression involved modeling the probability of a discrete outcome given an input variable (Dale, 2020). The binary logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. This model is of the form;

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}$$

$$= \frac{\exp(X\beta)}{1+\exp(X\beta)}$$

$$= \frac{\exp(X\beta)}{1+\exp(-X\beta)},$$

Where π is the probability that an observation is in a specified category of the binary Y variable, generally called the "success probability."

The likelihood for a binary logistic regression is given by:

$$L(\beta, y, X) = \prod_{i=1}^{n} \pi_i^{yk}(1-\pi_i)^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}\right)^{y_i} \left(\frac{1}{1+\exp(X_i\beta)}\right)^{1-y_i}$$

The log likelihood is obtained as;

$$l(\beta) = \sum_{i=1}^{n}[y_i log(\pi_i) + (1-y_i) log(1-\pi_i)]$$

$$= \sum_{i=1}^{n}[y_i X_i\beta + \log(1+\exp(X_i\beta))]$$

Since the maximum likelihood does not have a closed form solution in this case, iteratively reweighted least squares is used to find an estimate of the regression coefficients, $\hat{\beta}$.

2.3 Support Vector Machine Model

Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems (Busuttil, 2003). The idea of SVM is to create a line or a hyper plane which separates the data into classes. The SVM algorithm finds the points closest to the line from both the classes. These points are called support vectors. Then,the distance between the line and the support vectors is computed. This distance is called the margin. The goal is to maximize the margin. The hyper plane for which the margin is maximum is the optimal hyper plane (Figure 1). The SVM tries to make a decision boundary in such a way that the separation between the two classes is as wide as possible

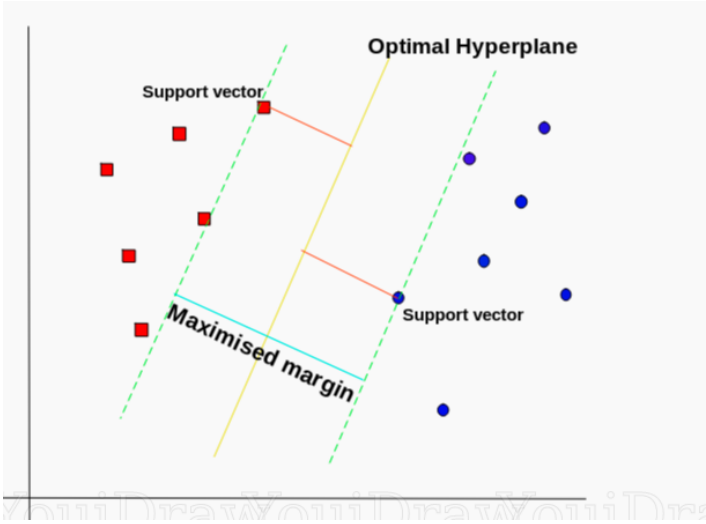

Figure 1: Optimal hyper plane in support vector machine model

Mathematically, the hyperplane equation is given as;

$$w^T x + b = 0$$

The margin lines are written as;

$w^T x + b = 1$ and $w^T x + b = -1$,

Depending on if the margine line is in the positive are or the negative area respectively

The distance between the two margins lines is given by;

$$w^T(x_2 - x_1) = 2 \Rightarrow x_2 - x_1 = \frac{2}{\|w^T\|}$$

The model aims to find the values of w and b that maximize the function;

$$(w^*, b^*) max \frac{2}{\|w^T\|} y_i \begin{cases} +1 \; where \; w^T x_i + b >= 1 \\ -1 \; where \; w^T x_i + b <= -1 \end{cases}$$

In a similar way, the above function can be expressed as;

$$(w^*, b^*) max \frac{2}{\|w^T\|} y_i * (w^T x_i + b_i) >= 1$$

The function can be expressed as a minimize function as;

$$(w^*, b^*) min \frac{\|w^T\|}{2} + C \sum_{i=1}^{n} \varepsilon_i$$

Where $C$the number of errors that will be misclassified is $\sum_{i=1}^{n} \varepsilon_i$ is the sum of errors

2.4 Models Evaluation and Comparison

The models evaluation and comparison was done using accuracy, precision, recall and F1 score. The definition of the above metrics is as below:

$$\text{Accuracy} = \frac{\text{True Positive+True Negative}}{\text{Total Population}}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True positive+false positive}} = \frac{\text{True Positive}}{\text{Total Predicted Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} = \frac{\text{True positive}}{\text{Actual Positive}}$$

$$F1 \text{ Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision+Recall}}$$

III. Results and Discussion

3.1 Data preprocessing and Variables Visualization

The first step was data cleaning. This ensured that columns are consistent. Visualization and summary statistics was an important step before fitting any model as it gave a glimpse of how the variables are associated with target variable. In this case stacked bar plots were used. They were used to check if the proportion of defaulters and non-defaulters was equal in different categories of a variable. From the graphs, it was shown that the proportion of defaulters and non-defaulters was different for the different credit history categories (Figure 2). This was also seen in the property area. From the categorical variables it was concluded that one of the best predictors was credit history. The values 1 represented defaulters and the value 0 represented non defaulters.

Figure 2: Distribution of defaulters and non-defaulters for different categorical predictor variables

For the numeric variables boxplot were used to visualize which distribution was different from the other. Non overlapping boxplot for defaulters and non-defaulters indicated that the mean or the median values in the two groups were significantly different. From this, it was be seen that it was unlikely that education and self-employment affected loan repayment and for this, these two variables could be dropped (Figure 3).

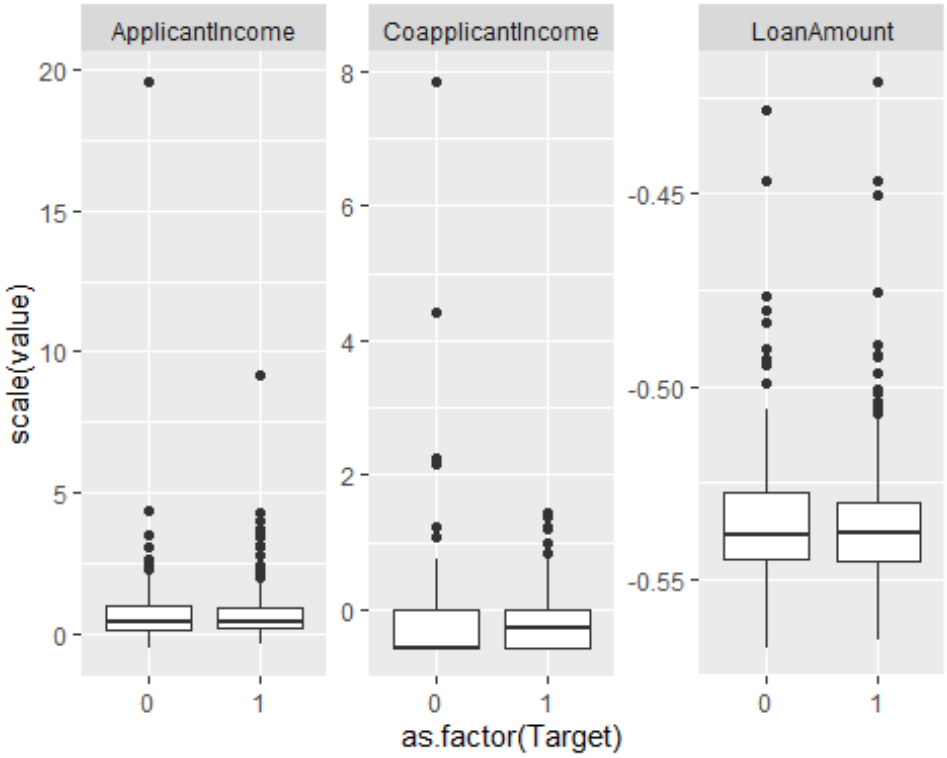

Figure 3: Distribution of defaulters and non-defaulters for different numeric predictor variables

Data pre-processing was done before models were fitted. This data preprocessing included; one-hot encoding for categorical variables with more than 2 levels, scaling variables and splitting of the data into test and train sets. In one-hot encoding for categorical variables with more than 2 levels, variables with more than two categories were converted into dummies variables. Scaling of variables was important since it led to faster convergence and since some algorithm used distanced to find decision boundary this meant that variables with big values had a big influence. Splitting of the data into test and train sets helped to evaluate the fitted model on data the models had never seen. The models were trained on the training set and tested using the test set.

3. 2 Binary Logistic Regression

Logistic regression was fit to predict the probability of an individual defaulting. The advantage of logistic regression was that the association between a predictor and response value could be seen and it also gave a probability. This was very important because a cutoff point could be set. For example, someone could be labelled as a defaulter if the predicted probability was more than 0.7. This increased precision but lowered recall. Using stepwise selection the fitted model was used to select the variables that best predicted credit default. The fitted logistic regression model is presented in Table 1.

Table1: Fitted Binary Logistic model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.2603698	15.599115	0.0807975	0.935603
Married	0.1178575	0.1737466	0.6783296	0.045628
Applicant Income	-0.126447	0.1404769	-0.9001263	0.036053
Loan_Amount_Term60	1.2508146	121.18821	0.0103213	0.991765
Loan_Amount_Term120	1.469752	145.65115	0.0100909	0.019488

Loan_Amount_Term180	0.6798044	0.2928181	2.3215931	0.020259
Loan_Amount_Term300	0.0956875	0.1632831	0.5860219	0.047608
Loan_Amount_Term360	0.6982726	0.2807777	2.4869232	0.012853
Credit_History	1.4407027	0.2264072	6.3633254	0
Property_AreaRural	-0.4157704	0.207827	-2.0005606	0.0454398
Property_AreaUrban	-0.4905839	0.208253	-2.3557107	0.0184873

The estimate column shows the log odds. Positive values means that the variable made it more likely for a person to repay their loan negative values means that the person is less likely to repay. The probability values of less than 0.05 meant that the variable under consideration was significant in predicting credit default.

3.2.1 Confusion Matrix for the Binary Logistic Regression Model

The confusion matrix was used to evaluate correctly classified cases. A perfect fit will have all values in the main diagonal while the entries of lower or upper triangular should be zeros. In this case we have 18 cases of false positives and 2 cases of false negatives (Table 2).

Table 2: Confusion matrix for the binary logistic regression model

	Non-defaulters	Defaulters
Non-defaulters	21	18
Defaulters	2	74

3.3 Support Vector Machine Model

The next step was to fit Support vector machine model. This was started by finding the best parameters using cross validation. The 10 fold cross validation was used where the train set was randomly split into 10 sets. In each case one of the sets was used as a validation or the test set while the other 9 were used to train the model.

3.3.1 Confusion Matrix for the Support Vector Machine Model

From the fitted support vector machine model there were 23 cases of false positives and 2 cases of false negatives (Table 3).

Table 3: Confusion matrix for the support vector machine model

	Non-defaulters	Defaulters
Non-defaulters	16	23
Defaulters	2	74

3.3.2 Validation Curves for the Fitted Support Vector Machine Model

A learning curve was used to find out if the fitted SVM suffered from high variance or bias. In this case the support vector machine model suffered from high bias (Figure 4). This meant that adding more data would solve accuracy problems.

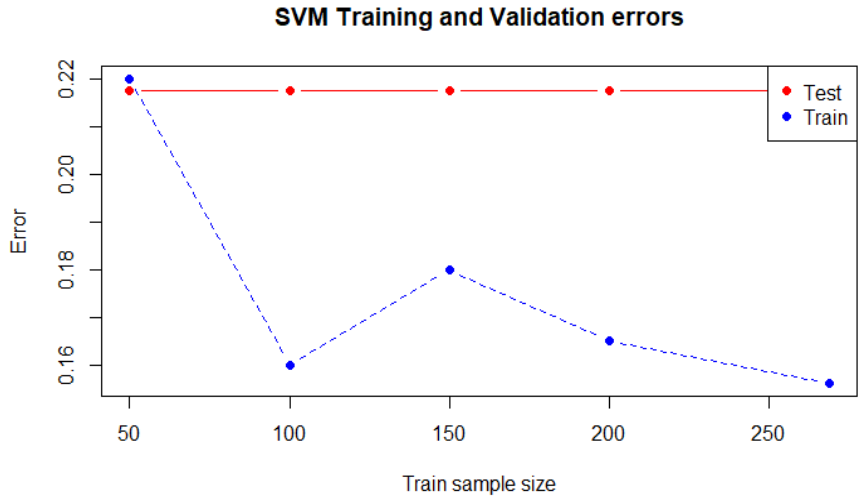

Figure 4: Validation curves for the fitted support vector machine model

3.4 Comparison of the Models for Predicting Credit Defaults

The performance of the fitted models was evaluated using accuracy, area under curve and F1 scores among other performance evaluation metrics. The accuracy of the binary logistic model was 0.83 with and F1 score was 0.89. F1 score was a very important evaluation metric where there are unbalanced classes. The area under curve for the binary logistic model was 0.74. Area under curve is important since it helps to know if the model suffers from high false negatives or false positives. A value of greater than 0.8 is normally desired. The accuracy for the fitted support vector machine model was 0.78 with and F1 score of 0.86 (Table 4). The area under the curve for the fitted support vector machine model was 0.692. This showed that the binary logistic regression performed slightly better than the support vector machine model in terms of accuracy, F1 score and area under the curve (Table 4).

Table 4: Performance metrics for the fitted logistic credit default models

	Binary Logistic Regression Model				Support Vector Machine Model			
Term	estimate	conf.low	conf.high	p.value	estimate	conf.low	conf.high	p.value
accuracy	0.826087	0.744264	0.890412	0.285385	0.782609	0.696036	0.854103	0.968532
kappa	0.568966	NA	NA	0.000796	0.441856	NA	NA	6.33E-05
sensitivity	0.804348	NA	NA	NA	0.762887	NA	NA	NA
specificity	0.913044	NA	NA	NA	0.888889	NA	NA	NA
pos_pred_value	0.973684	NA	NA	NA	0.973684	NA	NA	NA
neg_pred_value	0.538462	NA	NA	NA	0.410256	NA	NA	NA
precision	0.973684	NA	NA	NA	0.973684	NA	NA	NA
recall	0.804348	NA	NA	NA	0.762887	NA	NA	NA
f1	0.880952	NA	NA	NA	0.855491	NA	NA	NA
prevalence	0.8	NA	NA	NA	0.843478	NA	NA	NA
detection_rate	0.643478	NA	NA	NA	0.643478	NA	NA	NA
detection_prevalence	0.66087	NA	NA	NA	0.66087	NA	NA	NA
balanced_accuracy	0.858696	NA	NA	NA	0.825888	NA	NA	N

IV. CONCLUSION

In conclusion, the binary logistic regression and support vector machine models had accuracies of above 75% in predicting credit defaults. However, the binary logistic regression showed better prediction power when compared with the support vector machine model. The precision of the binary logistic regression increased with increased cut off point. From this study, it can therefore be recommended that binary logistic regression model can be used in predicting credit defaults. These prediction models should be integrated with a credit evaluation software where they can help credit officers decide if they will award a credit. In addition, there should be continued effort of evaluating if there are other algorithms such as XGBoost which uses boosting and bagging to find out whether they yield better classification accuracy than the ones considered for this study.

REFERENCES

[1] Bach, M. P., Zoroja, J., Jaković, B., & Šarlija, N. (2017, May). Selection of variables for credit risk data mining models: preliminary research. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1367-1372). IEEE.

[2] Brooks, J. R., & Levitin, A. J. (2020). Redesigning Education Finance: How Student Loans Outgrew the" Debt" Paradigm. *Geo. LJ*, *109*, 5.

[3] Busuttil, S. (2003). Support vector machines.

[4] Dale, E. B. (2020). *Introduction to Binary Logistic Regression and Propensity Score Analysis*. Working Paper. www. researchgate. net Accessed on 06/08.

[5] Edwards, J. (2019). What is predictive analytics? Transforming data into future insights. *CIO.[Online] Available at: https://www. cio. com/article/3273114/what-is-predictive-analyticstransforming-data-into-future-insights. html [Accessed 10 August 2019]*.

[6] Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2014). Modelling credit risk with scarce default data: on the suitability of cooperative bootstrapped strategies for small low-default portfolios. *Journal of the Operational Research Society*, *65*(3), 416-434.

[7] Ganong, P., & Noel, P. J. (2020). *Why do borrowers default on mortgages? A new method for causal attribution* (No. w27585). National Bureau of Economic Research.

[8] Geisser, S. (2016). *Predictive Inference*. Retrieved from *https://www. routledge.com /Predictive-Inference/Geisser/p/book/9780203742310.*

[9] Glennon, D., & Nigro, P. (2011). Evaluating the performance of static versus dynamic models of credit default: Evidence from long-term small business administration-guaranteed loans. *Journal of Credit Risk*, *7*(2), 3-35.

[10] Lunt, M. (2013). *Introduction to Statistical Modelling*: Linear Regression. Rheumatology, 54(7), 1137-1140.

[11] Masai, J. M. (2020). *Modelling Time to Default on Kenyan Bank Loans Using Non-parametric Models* (Doctoral dissertation, University of Nairobi).

[12] Sheskin, D. J. (2011). Parametric Versus Nonparametric Tests. *In International Encyclopedia of Statistical Science* (pp. 1051-1052). Springer, Berlin, Heidelberg.

[13] Wanjohi, S. M., Waititu, A. G., & Wanjoya, A. K. (2016). Modeling Loan Defaults in Kenya Banks as a Rare Event Using the Generalized Extreme Value Regression Model. *Science Journal of Applied Mathematics and Statistics*, *4*(6), 289-297.