ISYE7406 - Data Mining & Statistical Learning

CUSTOMER CHURN PREDICTION
A TELECOMMUNICATIONS CASE STUDY

April 25, 2021

Pablo González
Georgia Institute of Technology
pablogv@gatech.edu

Customer Churn Prediction

— Pablo González —

Abstract

Customer churn is a relevant and challenging problem common to most business and industries. As it is believed to be six to seven times more expensive to acquire a new customer than it is to keep a current one, understanding what characteristics or behaviors lead to a formerly loyal customer to end his relationship with a company is crucial for managing a sustainable business growth. In light of this, the following paper aims to study the use of different data mining and statistical learning methods to effectively predict customer churn in a telecommunications company. For this purpose, five different models were trained and their parameters were tuned to adjust for imbalanced data. Subsequently, relevant evaluation metrics were calculated and models were evaluated accordingly. Finally, we conclude that through the use of a class-weights adjusted random forest we can obtain cost-effective actionable insights to significantly reduce customer churn.

Introduction.

Customer churn, also known as customer attrition, defection or turnover, is a relevant and challenging problem that most businesses have to handle as even the largest, most stable companies suffer from this situation and understanding what characteristics or behaviors lead to a formerly loyal customer to end his relationship with a company is crucial for managing a sustainable business growth. Because the costs of acquiring new customers can be higher than the ones of retaining loyal ones, we believe that it is only wise to attempt to better understand this problem and target likely to churn customers with specific campaigns to raise their perceived value, as well as their satisfaction and retain them.

On the other hand, as we know, "Data Mining & Statistical learning" refers to a framework of statistical models used for both inference and prediction, thus including also the ability of predicting the class of a given data point. For example, whether a certain customer is likely to defect.

This paper aims to respond to the following scientific question: Can we predict customer churn of a telecommunications company and, if so, how to better predict which customer is likely to churn? In order to do so, we propose the use of data mining & statistical learning methods to model whether a certain customer is likely to turnover and be able to target him with a specific marketing campaign.

Problem Statement and Data Sources

As previously indicated, this paper studies how to predict customer churn in a telecommunications company. Thus, in order to better understand the business problem and our research question, it is important to be aware of some of the most important factors tha will affect our decision when selecting our final predictive model.

For illustrative purposes, let us consider the case of a regular telecommunications company that is experiencing higher churn rates than they would like to. We might not know much about the current financial situation of this company, their competitors nor the market situation in which they are operating. However, we do know that:

1. Their churn rate is large enough to represent a problem and yet small enough for our data to be imbalanced, as they are still operating in the market.
2. Their churn rate is impending them from meeting their goals and they need to design a marketing strategy that will allow them to rettain valuable customers that might be considering defecting to a competitor.
3. Regardless from the financial situation of this company, they will require this stratetgy to operate within a certain budget constraint and as cost effectively as possible.

Thus, in order to solve this problem, we must not only generate a model that has a low misclassification error, but we also need it to be certain when making a decision (i.e have well defined decision boundaries), as well as sensitive and cost-effective whenever deciding which customer is likely to defect, so as to reduce our operation costs by refraining from including other customers in our strategy.

As a result, in this paper, we will be working with the *"Telco Customer Churn"* data set. This data set contains information about a fictional telecommunications company that provided home phone and Internet services to 7043 customers in California in 2019Q3. Thus, our data is composed of this same number of observations and 21 variables with valuable information including contracted services, demographic characteristics and whether the customer in question hast churned or not.

In lights of this, in the following sections we will determine which variables are more likely useful for this classification problem and quantitatively compare the performance of 5 different statistical methods in terms of both error rate (1-accuracy), area under the ROC curve (AUC) and sensitivity, also known as true positive rate. Since our aim is of predictive nature, our decisions regarding variable selection will be taken accordingly. Additionally, no coefficients will be presented, as a significant amount of the employed models do not allow this type of interpretation.

Exploratory Data Analysis

When first analyzing our data set, we could notice that 11 observations had a missing value. Upon research it was discovered that all of these missing values corresponded to a newly acquired customer and that all of these new customers missed information regarding *TotalCharges*. Thus, as they did not provide any useful information, these observations were deleted, resulting in our data set only containing information about customers that have stayed at least 1 month in the company. Additionally, no outliers were found according to the inter-quartile range method.

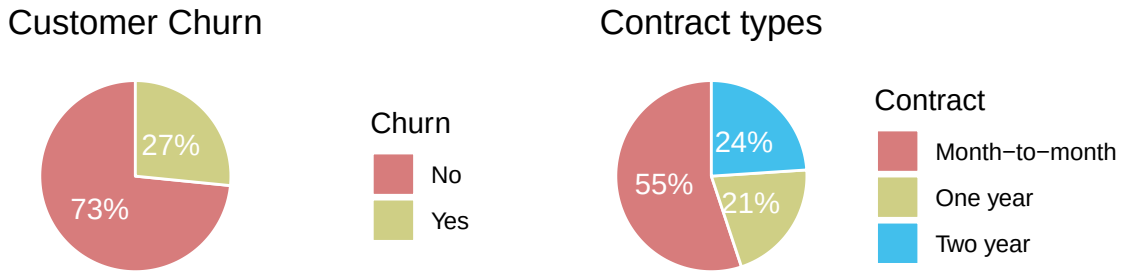
Our processed data is composed of 7032 observations (customers) and 20 variables. For each row, the first column contains a binary variable denoting whether the customer in question has defected or not and the remaining 19 variables represent certain characteristics that may provide insightful information to train our models and predict this state.

In order to better understand our potential predictors and learn which information might be more useful for our analysis, we calculated the information gain of each variable through entropy and arranged them in descending order, as shown in Figure 1 on *Appendix A*. Thus, when analyzing these results, we can notice that *MultipleLines*, *PhoneServices*, as well *gender* have close to no predictive power, whilst *Contract* can be considered a strong predictor. It is, however, worth noting that these variables are not therefore dropped, but, in accordance with modern data mining literature, we let our algorithms auto-select them whenever possible.

Additionally, we can calculate the correlation between our potential predictors, as shown in Figure 2 on *Appendix A*, noticing that additional services, as well as tenure are, as expected, highly correlated with charges, which could cause multicollinearity. However, we do realize that, whilst it is important to acknowledge these variables having a strong correlations between each other, thus violating the non-collinearity assumption of the logistic regression, as we know, multicollinearity does not affect the accuracy of predictive models and removing an important predictor can generate omitted variable bias, thus affecting our predictive goal.

After having analyzed the information given by our predictors, we would like to further analyze both our response variable and potential predictors. To illustrate, Figure 3 shows the distribution of both our response variable and the predictor with the highest information gain. It is important to note that the variable *Churn* is not equally distributed. Additionally, we can also notice that "Month-to-Month" contracts account for more than 50% of the contracts in our data set, whilst only a 21% are annual contracts.

Figure 3: Most relevant Categorical variables



Additionally, Table 1 shows some basic summary statistics for our numeric predictors.

Table 1: Summary statistics

	Min	**Mean**	**Median**	**Max**	**Var**	**Kurtosis**
tenure	1.00	32.422	29.000	72.00	602.470	-1.388
MonthlyCharges	18.25	64.798	70.350	118.75	905.166	-1.257
TotalCharges	18.80	2283.300	1397.475	8684.80	5138252.407	-0.233

Regarding the Kurtosis statistic, it is important to note that this value measures the heaviness of the tails of a distribution relative to a normal distribution, with a Kurtosis value of 3. This means that variables with a higher Kurtosis are more prone to presenting outliers and, in contrast, variables with a negative kurtosis tend to have a flatter distribution.

Finally, we decided to split the original data set into disjoint training and testing data sets, so that we can better evaluate and compare different models. By doing so, we will train our models using 90% from our original data, thus allowing for a better model generation, and keep the remaining 10% for testing the accuracy of these models.

Proposed Methodology

As previously stated, in this report we will compare the performance of 5 different statistical methods for our classification problem. Thus, in order to achieve this goal, we run the following models:

1. Naive Bayes (NB): This probabilistic classifier assumes independence of the predictors and uses the Bayes rule to compute the conditional posterior probabilities of our response variable. For this analysis we use a gaussian naive bayes classifier, which implies that we also assume a conditional gaussian likelihood for the predictors. Thus, although still working, this classifier tends to present problems when facing categorical independent variables.

2. Logistic Regression (LR): This classifier is an extension of the linear regression where the target variable is of categorical nature and is therefore predicted using the logit of the odds as the response variable. This method is commonly used for binary classification problems, as it makes no assumptions regarding the distribution of the target classes. However, it can become unstable when there is a clear separation between classes. Finally, we use a cutoff of 50% when doing the classification and variables are standardized and selected based on Akaike information criterion (AIC).

3. Random Forest (RF): Similar to the previous method, Random Forest can also be used both for regression and classification problems. In this case, this ensemble classifier operates by constructing 300 decision trees and taking the majority vote of these trees for prediction. As we know, ensemble learners generally outperform all their constituents, however, data characteristics can still affect their performance. Therefore, the minimum size of terminal nodes was set to 3 and 2 variables were randomly sampled as candidates at each split. Additionally, class weights were slightly adjusted in order to favor customers likely to churn by a margin of 10%.

4. Boosting (BST): Boosting is also an ensemble learner. This supervised algorithm is characterized by being able to convert several weak learners to a strong one, thus allowing for bias an variance reduction. The optimal number of boosting iterations is estimated through a 10-fold cross-validation, learning rate is set to 1%, cutoff value is set to 45% and one thousand trees are generated per iteration.

5. Support Vector Machines (SVM): Similar to previous methods, this supervised algorithm is also useful for both regression and classification problems, both linear and non-linear. Support vector machines work by creating an hyperplane that separates data into different classes with minimum error. For this problem, a linear kernel with standardized predictors and soft margins with inverse class weights are used in order to favor the prediction of positive Churn values.

In the following section, we will analyze the results of the above presented models. It is important to note that no information regarding coefficients was given due to the fact that a significant amount of our models do not generate this type of values and also because our aim is of a predictive nature and we have already determined that most of our independent variables have predictive power over our response variable.

Analysis and Results

As previously stated, in this section we will test our 5 models on unseen test data and discuss their performance using our previously selected evaluation measures. For this purpose, we would like to introduce these measures once again and explain how they relate to our research question.

- Misclassification error: As its name indicates, this rate represents the percentage of times in which our model fails when predicting whether a certain customer will churn or not. Thus, a better performance is indicated by a smaller value.
- Area under the roc curve (AUC): This value indicates how well a model is capable of correctly distinguishing between the given classes and, therefore, a better performance is indicated by a larger value.
- Sensitivity: Also known as true positive rate, this measure tells us the percentage of churning customers that are accurately predicted by our model and, therefore, a better performance is indicated by a larger value.

Note: Whilst we understand that the specificity would represent a better indicator of misspent resources, we consider that the misclassification error gives us a more general insight regarding the models' prediction capability, while, in conjunction to the sensitivity, also allowing for the attainment of this same information.

Now that we understand our evaluation measures, Table 2 reports all three of our evaluation metrics for each model.

Table 2: Evaluation measures

	NB	LR	RF	BST	SVM
Misclassification	0.28165	0.19488	0.27596	0.20626	0.34993
AUC	0.80874	0.82399	0.83242	0.82887	0.81383
Sensitivity	0.79503	0.50311	0.75776	0.52174	0.86335

When analyzing these results, we can notice that both the naive Bayes and random forest classifiers seem to consistently provide good results for all of our evaluation metrics.

In fact, the misclassification error indicates us that the best model for this testing data set in terms of is given by the Logistic regression classifier with a testing error of approximately 19.5%, followed by the Boosting algorithm with a testing error of approximately 20.6% and then the random forest classifier.

Additionally, when analyzing the AUC, we can notice that the Random Forest classifier is now the best performing algorithm and that the Boosting classifier comes again in second place.

Furthermore, when considering the sensitivity for each model, we can notice that the Support Vector Machines classifier performs the best in terms of this measure. However, as it did not perform as well in terms of misclassification error, now the balance tilts between the naive bayes classifier and the random forest, which is more likely to be our final choice due to its stable good results in each evaluation measure.

Normally, this analysis should be sufficient if we one had a large data set. However, since our data set is relatively small, we decided to use Monte Carlo cross-validations to further assess the robustness of each method. Table 3 reports the summary statistics for our testing error after performing a hundred repetitions.

Table 3: Summary statistics for the sensitivity

	Min	**Mean**	**Median**	**Max**	**Var**	**Kurtosis**
NB	0.71591	0.80138	0.80541	0.85311	0.00082	0.13177
LR	0.46561	0.54879	0.54959	0.66169	0.00131	0.01171
RF	0.70526	0.78681	0.78995	0.85864	0.00116	-0.46778
BST	0.46512	0.57738	0.57821	0.68657	0.00134	0.69013
SVM	0.78836	0.86914	0.86895	0.92268	0.00059	0.84655

When first analyzing the sensitivity of our models, we can notice that both the logistic regression and boosting classifiers have a significantly low sensitivity compared to the other models. Therefore, since our first priority is to detect customers that will churn, we decide that they do not fit our targets and stop considering them in the analysis. This problem can be partially solved via downSampling, but the use of this statistical method can cause information loss and does, in fact, rise the misclassification error as well.

Table 4: Summary statistics for the misclassification error

	Min	**Mean**	**Median**	**Max**	**Var**	**Kurtosis**
NB	0.22475	0.27799	0.27738	0.31721	0.00026	0.66838
LR	0.16358	0.19771	0.19844	0.25462	0.00021	1.18524
RF	0.21195	0.24499	0.24467	0.28734	0.00023	-0.29513
BST	0.16785	0.19862	0.19986	0.24324	0.00021	0.34610
SVM	0.29445	0.33514	0.33428	0.36842	0.00031	-0.66228

Additionally, when analyzing the misclassification errors, we can notice that they have now slightly increased. However, the random forest classifier is still the better performing algorithm among the remaining models. We can also notice that whilst the Support vector machines classifier outperformed its counterparts in terms of sensitivity, it does have a significantly high misclassification error of approximately 33.5% and therefore might not be the best choice if facing strict budget constraints.

Furthermore, when analyzing the summary statistics for the area under the curve (AUC) given by Table 5, we can notice that the random forest classifier does also significantly outperform the remaining models, as it has a AUC value of approximately 84% with a variance of only 0.023%.

Table 5: Summary statistics for the AUC

	Min	**Mean**	**Median**	**Max**	**Var**	**Kurtosis**
NB	0.76834	0.81584	0.81723	0.85469	0.00030	0.35572
LR	0.79392	0.84324	0.84300	0.87132	0.00022	0.57314
RF	0.79203	0.84011	0.83888	0.87282	0.00023	0.50992
BST	0.79320	0.84603	0.84354	0.87847	0.00025	0.45556
SVM	0.17989	0.82505	0.83157	0.86251	0.00450	82.58081

After having performed the previous analysis, we noticed that the random forest classifier outperforms all other algorithms when evaluated on all of our performance measures. And, whilst it is not the best performing model in terms of sensitivity, it still achieves good results by predicting, on average, approximately 78.7% of customers turning over.

Additionally, it is also important to notice that the support vector machines classifier is the best model if we only care about sensitivity and not so much about our budget constraints, due to subsequent costs such as the ones of the marketing campaign.

Finally, taking these results into account, in the following section we would like to give a general summary of our work and conclude this report with some general remarks (conclusions) regarding the performance of our 5 different classifiers.

Conclusions

As previously indicated, our analysis shows that the random forest classifier outperforms all other algorithms when evaluated on all of our performance measures. To confirm the statistical significance of our hypotheses, and since we do not know whether our data is normally distributed, we conduct both a T-test and a Wilcox-test over the previously presented performance measures. Analysing these results, which can be found on *Appendix B*, and in lights of our previous statements, we can notice that, in fact, they are statistically significant, since we can reject the null hypothesis of them having a similar performance, to the 95% level.

The previous tests prove that our conclusions regarding the best classifier in terms of both misclassification error, AUC and sensitivity were statistically significant. As a result of this, regarding our recommended model to predict *Churn*, we argue that the Random Forest classifier is the best model for our declared objectives.

In conclusion, after the conducted research, we argue that it is not only posible to predict customer churn, but, using the random forest classifier, we can obtain actionable insights to effectively reduce customer churn in our company.

Finally, regarding future improvements, we realize that our models are far from perfect and much work can still be done, specially in terms of outlier detection and handling of imbalanced data. Thus, it would be interesting to see future studies in these subjects, as well as studies regarding selection of interaction variables.

Lessons learned

The decisive factor that made me get interested into the world of analytics and data science was through a course called "Statistical Learning". Therefore, after reading that this course was being added to the program, I immediately decided that this was a course that I needed to participate in.

This class has allowed me to learn more about this topic and renew my interest in the area. I have realized that there is much to learn and I also need to improve my mathematical skills in order to fully understand some algorithms and make some relevant proofs.

Additionally, regarding this final project. This has been the first time I have ever truly noticed the problems of working with imbalanced data. For this purpose, both over and under sampling were initially considered and their benefits, especially in terms of computing costs for the under-sampling method, were analyzed. However, after noticing that these methods generated a higher misclassification error, I decided to adapt the in-class learned algorithms to be able to handle this type of data, obtaining some specially good results in the random forest classifier.

I believe that the course was well made and do not have many complaints regarding the peer evaluations. However, I would recommend to making it a requirement to give an X-word minimum feedback when doing these assessments.

Thank you very much for the experience.

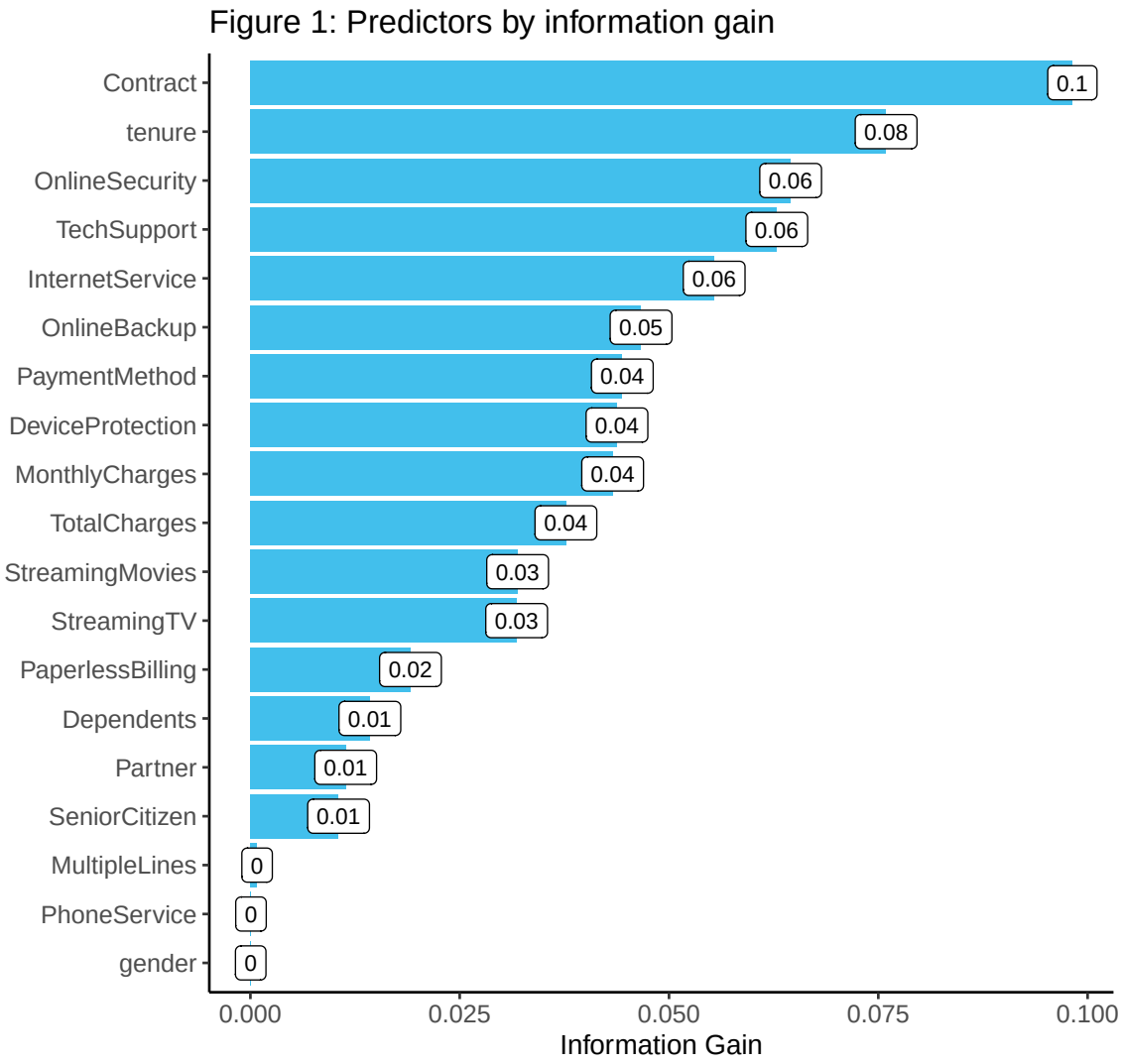
References

- This paper was done using the "Telco Customer Churn" data set from Kaggle's repository. Available from: https://www.kaggle.com/blastchar/telco-customer-churn.
- All data analysis was done on a Windows 10 Laptop with Intel i7-10510 CPU 1.80GHz using R 4.0.3 (R Core Team, 2020).

Appendix

A. Analysis of predictors.

Information gain:

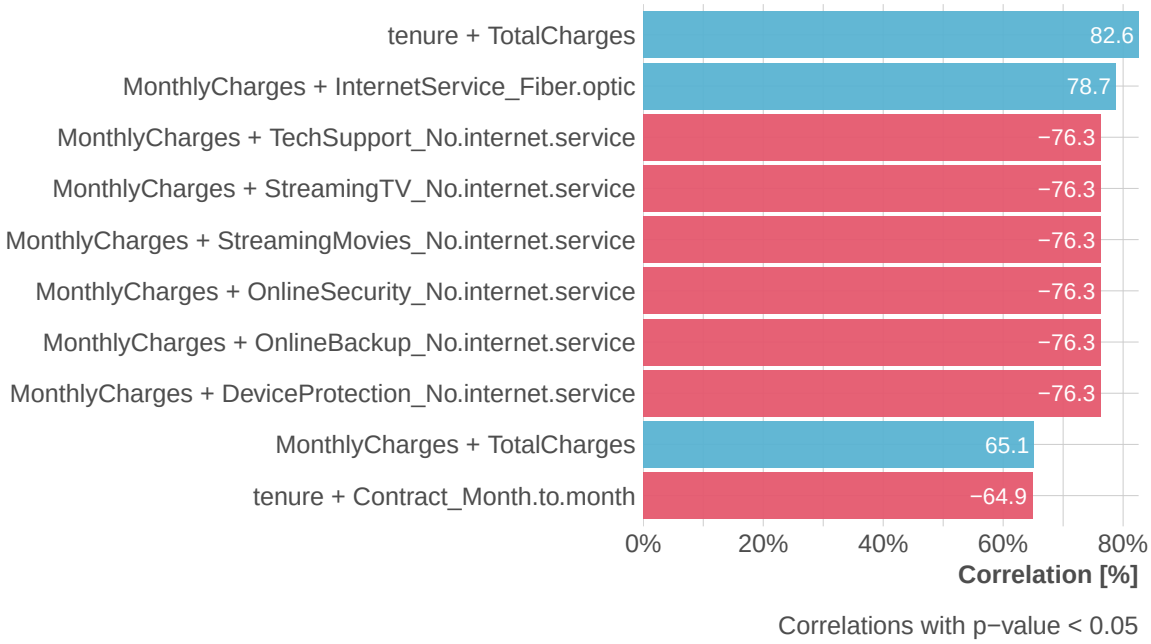


Predictors with the highest correlation values:

Figure 2

Ranked Cross−Correlations

10 most relevant



B. Significance of the obtained results

As previously indicated, in this section we assess the statistical significance of the differences between performance of our models, from which our final conclusions were drawn. For this purpose, we conduct both a T-test and a Wilcox-test over the presented performance measures, obtaining the following results (All values have been approximated to the fifth decimal place).

Table 6: A misclassification error comparison of Random Forest against other methods

	NB	LR	BST	SVM
T-test	0	0	0	0
W-test	0	0	0	0

Table 7: An AUC comparison of Random Forest against other methods

	NB	LR	BST	SVM
T-test	0	0	0	0.019
W-test	0	0	0	0.000

Table 8: A sensitivity comparison of Random Forest against other methods

	NB	LR	BST	SVM
T-test	0	0	0	0
W-test	0	0	0	0

When Analyzing these results, we can notice that all previous statements are statistically significant to the 95% level.