Analysis of Supervised Machine Learning Techniques on 311 Cases, Fire Dispatch Incidents, and Loan Defaults

Donovan Michael Jeremiah
School Of Computing
National College Of Ireland
Dublin, Ireland
x18181562@student.ncirl.ie

Abstract— **Government services such as the 311 non-emergency service in San Francisco are regularly monitored to ensure that cases are attended to and resolved promptly. The ability to predict the resolution time of the 311 cases would be very helpful to the service authorities in evaluating which areas in public service need improvement. San Francisco's public dataset of over 3.8 million 311 cases have been used to predict the resolution time of a case using certain attributes of a 311 case.**

A City's Fire Department is usually the main respondent to any major accidents in the city and well equipped to handle the situation. This makes it crucial for them to reach the incident location as quick as possible. The response times of the fire department or any emergency service must be critically observed and evaluated to maintain quick responses. The New York City Fire Department's (NYFD) dataset of over 3.3 million incidences has been used to predict the response time of the fire department to the scene of the incident.

The usage of machine learning techniques in predicting loan defaulters has always been used by banks to mitigate the risks associate in offering a loan. This would benefit existing customers as well. For this, the dataset of vehicle loan applicants from L&T has been used to predict if the loan applicant will default in their first loan repayment installment.

Keywords – **311 cases, fire dispatch incidents, loan defaults, machine learning, Naïve Bayes Classification, KNN Classification, Multiple Linear Regression, Random Forest Regression, Logistic Regression.**

I. Introduction

A. Motivation and Description

Due to the recent water shortage crisis in the city I grew up in, I know first-hand how government services such as the City's water department are responsible for the timely delivery of drinking water to the neighbourhoods. India's urban cities alone are home to over 43% of the country's 1.3 billion population.[1] San Francisco is a city with a large population as Chennai. Analysing its 311 cases would give us insight on how metropolitan cities fair in terms of resolution time of the 311 tickets raised. In this report, the KKN and Naïve Bayes classifiers will be trained using a dataset of 311 cases from San Francisco in 2019 to predict the future response times of NYC's fire services department. The Root Mean Squared Error (RMSE) and the R-Squared value were among the evaluation methods used. Actual vs fitted plots were also used to visualize the efficiency of the models.

Response times for fire departments are extremely useful in analysing the areas of improvement in firefighter reaction times. Survival rates are affected more when response time targets are not met.[2] In this report, we will see how we can use it to train a model to predict the future response times of NYC's fire services department by analysing the dataset of NYFD's fire dispatch incidents and their measurements such as alarm times, time of day, area of the incident, etc. Among the evaluation methods used were the confusion matrix and the ROC curve.

The benefit of accurately predicting loan defaulters would result in lesser risk costs to banks and as a result, have an impact on the loan interest rates offered to regular-paying customers.[3] In this report, we use the L&T dataset of vehicle loan applications to predict if an applicant will default in their first loan repayment instalment. The Logistic Regression model will be trained with this dataset and evaluated using the confusion matrix among other evaluation methods.

B. Research Questions

- What is the predicted resolution time class for a 311 case?
- What is the predicted response time for the first fire department unit to reach the incident?
- Will a loan applicant default the first loan repayment instalment?

This paper is organized as follows, in which section II highlights the related work in the respective domains, section III demonstrates the methodology used to perform the machine learning predictions, section IV contains the results and discussions. Section V contains the conclusions and future work, followed by the references in section VI, and the footnotes in section VII.

II. Related Work

The papers discussed in this section is a comprehensive study of the related work with papers published on the study of the duration of services cases and the methods used to predict them. In addition, nuanced aspects related to machine learning techniques used later in this report will be attributed to in this section. J. Goderie *et al.* [1] in their paper discuss the KNN (K Nearest Neighbors) algorithm, which is a supervised classification algorithm, and its use in predicting the time duration bin in which a Stack Overflow question will be answered with the highest user-voted answer. It is also observed that there is no significant increase in performance when the value of k is increased. This will be taken into consideration later in the evaluation section of this paper when KNN is used to classify resolution times of 311 cases. Their model was found to better predict user-accepted answers since they usually had shorter times. An approximate of 30-35% of posts were predicted correctly after using a 10-fold cross-validation. S. Boyles *et al.* [2] demonstrate the use of the Naïve Bayes Classification method to predict time durations of incidents occurring in traffic which could include road accidents, traffic jams, protests, etc. They claim that the method is suited best for the chosen dataset. The binning method chosen here served as a reference to the binning of the dependent variable, resolution time, in 311 cases. This paper [3] by Y. Lin *et al.* takes a different approach to the papers mentioned above and predicts the demand of logistics delivery, rather than the time of delivery, citing the increasing customer need for on-time delivery as motivation. It seeks to predict the delivery demand in each sub-region based on spatial-temporal data and warns about the existence of too many factors in the predictor variables and their relative distribution in real-time data which hints to an unbalanced dataset. We face this problem in the 311 cases when we try to predict the resolution time of cases closed after a year, which is rare. They then illustrate the workings of the LSTM (Long Short-term Memory) neural network to predict demand locations and demand times. In their paper [4], W. Zhou *et al.* discuss KNN and its use as a recommendation technique to recommend resolutions for incoming tickets in the IT Service domain. After the KNN model is trained to run using just the other ticket level attributes and identifying the need for more accuracy, the model is improved to analyze the comments of a raised ticket. This prevents true negative tickets from being incorrectly recommended by the KNN algorithm as a false ticket, by using a penalty driven approach in such cases. To improve the base model several variations of KNN were constructed such as WKNN (Weighted KNN) and an LDA (Latent Dirichlet Allocation) flavored KNN. It is concluded that using a penalty scheme on the KNN used for the 311 cases would not be possible as there are no attributes that indicate a 311 case is falsely reported. H. Xu *et al.* [5] describes a statistically inclined technique called the Hazard Duration Model which uses a Hazard function to model a prediction algorithm used to predict durations that focus on end-of-duration specific time series. One of the function's main advantages is that it performs well due to its efficiency with risk-based analysis. Here, using the STATA statistical analysis software, the Kaplan Meier plots are used as a means of evaluation. Ryan Eshleman *et al.* [6] demonstrate the various analysis performed such as sentiment analysis on 311 complaints from the City of San Francisco. Using spatio-temporal analysis, they also investigate the emotional sentiment of tweets from twitter across major metropolitan cities across the globe in relation to the 311 cases reported. This counterintuitively results in the indication of a positive correlation between the sentiment of a city and the reported cases 311. The more 311 cases were reported in a city during a time period, the more positive was the sentiment of the city was during the same time. Techniques such as a Multinomial Naïve Bayes Classifier along with NLP accompanied with 10-fold cross-validation for sentiment analysis is used in the paper. In their paper [7], L. Ren *et al.* illustrate a more robust and improved SMV classification technique known as Binary Tree SVM Classification which works on a Bayesian probabilistic method of prediction. Later in this report, the reasons for rejecting SVM to predict the resolution time of 311 cases will be supported by in L. Ren *et al.*'s paper. A multiclass SVM classification algorithm called Bayesian-based Binary Tree SMV is examined and evaluated.

S. Deshpande *et al.* [9] uses time-series data and system workload information such as I/O block sizes, CPU usage metrics, etc., of enterprise-grade storage systems to train a Random Forest Regression model to predict the response time of storage systems in real-time continuously. Here the common Pearson's correlation coefficient is used to check variable significance. Metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used as evaluation metrics since the model used is a regression. It is worth noting that later in this paper, the RMSE value will be used to compare two different regression models fitted on the Fire Incident Dispatch dataset. In this paper [8], S. Singh *et al.* proposes a technique to predict the order lead time in a manufacturing organization that follows a JIT (Just-In-Time) manufacturing setup. The KNN Regression is used along with evaluation methods such as Mean Squared Error (MSE) and RMSE. Correlation coefficients such as Pearson, Kendall, and Spearman have been used to check the variable significance. J. J. Rebollo *et al.* [10] use a Random Forest Regression to predict delays in the departure of more than 60 minutes almost 2-24 hours in the future. A 10-fold cross-validation scheme is used. A Random Forest Regression was chosen over linear regression and even neural networks. Parametric ANOVA tests such as the Kruskal-Wallis test were performed to analyse the categorical explanatory variables, and the low p-values obtained proved the high significance of the explanatory variables. Error distribution plots were also examined for patterns. It is found that the average mean test error increases only by 7.4 minutes as the prediction goes from 2 hours to 24 hours in the future, which is considered to be incredibly reliable and useful. C. Huang *et al.* [12] use a classification method to predict the response time to assist the chatbots in providing a more human-like experience to the user. Response time in seconds is binned into 4 intervals from 0 up to 4 minutes, after which the chatbot will cease to continue the conversation. Again, we see the use of the LSTM neural network, as in Y. Lin *et al.*'s paper, along with the use of the Convolutional Neural Network (CNN). After appropriate encoding of explanatory variables, a 7:2:1 ratio is used for training, validation, and testing subsets of the dataset. This improved the accuracy of the CNN model to 48.6%, and the LSTM model to 50.1%, and hence, the LSTM is chosen as the more accurate model. S. Handley *et al.* [11] use a KNN prediction model to predict the travel time from the data on the City of San Diego. They claim that the value of k was chosen as 3 since a higher value did not aid in a better predictive model. Normalization was performed on the appropriate predictor variables to prevent either domination by any high magnitude predictors. A 10-fold cross-validation was also performed, but this time, as a method of feature selection. Their analysis found that the time of day predictor variable was more useful only when predicting longer trips, which is intuitive. Short trips could be made at any time, but longer trips were more likely to be carried out during the day in the presence of sunlight for clear vision while driving. B. Van de Vyvere *et al.* [14] demonstrate the prediction of traffic light phases to aid in route planning and duration in the City of Antwerp. Like in the above papers, MAE is also used here as a validation metric. Predictor variables were grouped as weekend or weekday, the period of the day, etc. as traffic varies between these times thus alter the traffic light phases as well. 10-fold cross-validation is performed and it is observed that as traffic light phases increase the MAE of the predictions also increase making it inaccurate to predict longer traffic light phases. Y. He *et al.* [13] designs a model to predict delays in travel time due to non-recurrent traffic events. They claim that the proposed neural network in the paper outperforms machine learning methods like multivariate decision trees. They train it with traffic and incident datasets of Lyon, France for 5 months. Their proposed model known as the Piecewise Affine Incident Model works on a combination of nonlinear regression models that are constrained and unconstrained. It is interesting to note their use of RMSE as a tool to filter data for the training set to improve accuracy. Incidents with an RMSE value of 6 or more were removed. A Univariate Decision Tree (UVDT) from SPSS and a Multivariate Decision Tree (MVDT) using the R package mvpart is compared to a neural network called the Multilayer Perceptron (MLP) method from SPSS. An 80:20 training to test ratio is used to split the dataset. After the Mean Absolute Error (MAE) is used to plot a distribution of it across the 3 different models, MVDT and MLP have been shown to outperform UVDT, while MLP achieves the smallest prediction error.

R. Kumar *et al.* [15] demonstrates the use of Logistic Regression in predicting the Clickthrough Rate (CTR) of online advertisements from a 25GB advertisement dataset using predictor variables such as position etc. Min-max normalizations were performed to scale down the predictor variables. After 10-fold cross-validation accuracy of about 90% is obtained using the logistic regression model. E. Afatmirni *et al.* [16], in their paper, attempt to predict the instance of Refibrillation, the recurrence of Ventricular Fibrillation (VF), that poses a significant problem to Emergency Medical Services (EMS) personnel during cardiac resuscitation such as Cardio Pulmonary Resuscitation (CPR). The proposed classification method which is a Linear Discriminant Analysis (LDA) based classifier achieved an accuracy of about 75-76.5% with a p-value of 0.0028. Despite having a small dataset to train, reasonably good accuracies were obtained due to the usage of the Leave-One-Out (LOOM). Y. Sayjadah *et al.* [17] conclude that their usage of the random forest as a classification method, in comparison with others such as decision trees and logistic regression to predict credit card defaulters, proves to be the best and results in an accuracy of 82%. Here, the Correlation-based Feature Selection (CFS) is utilized. A 70:30 split of the train to test data is used. The Area Under Curve (AUC) for the random forest was 77%. S. H. Ebenuwa *et al.* [18] suggest a method for overcoming the shortcomings of a class imbalance in classification prediction. An attribute selection technique called variance ranking is proposed in this paper, and its performance is compared with the Pearson's correlation and the information gain technique. SVM, logistic regression, and decision tree were used to evaluate the performance of this proposed attribute selection method to mitigate the adverse effects of class imbalance in your predictor variables. This paper also points to the Synthetic Minority Oversampling Technique (SMOTE) which is commonly used but indicates that it introduces a misclassification cost which hints to the motivation of this proposed method of attribute selection described in this paper. Note that an under-sampling method is used below in this report on the loan default prediction dataset. However, the variance ranking method performs well only when the target is numeric (continuous or discrete). It does not perform well with multiclass attributes. On the subject of multiclass attributes, it is useful to note that for the 311 cases resolution prediction, no reliable multiclass sampling method was found to aid in over/under-sampling the target variable and hence no sampling technique was used. T. S. Brisimi *et al.* [19] use kernel and sparse SVMs, logistic regression, and random forest to predict the hospitalizations due to heart diseases and diabetics. K-LRT, a likelihood ratio test-based method that is based on the Naïve Bayes classifier, is used fir interpretability of the results as it indicates k features which are the more significant. G. Sudhamathy et al. [20] use a decision tree classification method to predict credit defaulters. KNN is used to impute missing values, while SMOTE, an over-sampling method, is used to balance the binary class variables in the dataset. The correlation coefficient is used to remove insignificant predictor variables.

III. METHODOLOGY

The KDD (Knowledge Discovery in Databases) approach which was mentioned by Fayyad *et al*. (1996) has been used as the method for this data mining and machine learning project. This method details a flow of steps that need to be followed to eventually arrive at the insights that can be drawn from a database. These steps can be seen in Fig. 1. The rest of this academic paper will describe how the steps of the KDD methodology were used to make predictions and draw insights from the datasets used.

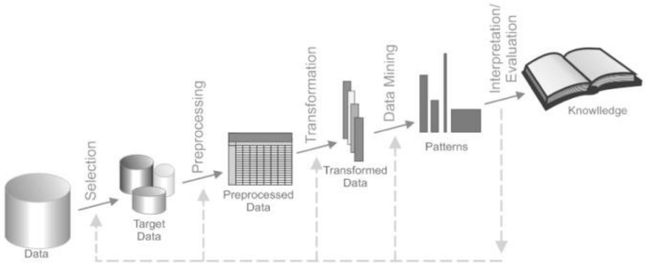


Fig. 1: KDD Life Cycle[7]

A. *Data Collection*

- The data used for the prediction of the resolution time of 311 cases were taken from the '311 Cases' dataset[4] from San Francisco's website of public data, DataSF. It has the data of 3.94 million 311 service cases with 20 attributes for each case.
- The data used for predicting the response times of fire department services were taken from the 'Fire Incident Dispatch Data' dataset[5] of NYFD from NYC Open Data, NYC's webpage of public data. It contains the data of 3.38 million fire service incidents in the city of New York. There are 29 attributes for each incident.
- The data used for the prediction of vehicle loan defaulters were obtained from Kaggle's 'L&T Vehicle Loan Default Prediction' dataset[6]. It contains the data of 230K unique loan applications with 41 attributes for each application.

Note: Refer footnotes for the links to the datasets.

B. *Data Pre-Processing*

The data datasets were large in size and filled with inconsistency issues such as missing values, outliers, etc. The below sections outline how the issues were dealt with.

1) Filtering Data: The '311 cases' and 'Fire Incident' datasets were filtered through and only the data from 2019 was taken for them resulting in around 500K rows. Furthermore, random sampling was done to reduce the size of the dataset to almost half

2) Handling Missing Values: Due to the abundance of observations in each dataset, and the insignificance of the predictor variables, those with missing values were removed.

3) Correlation Matrix: Attributes that were found to not contribute significantly to the accuracy of the predictive models were removed from the dataset during the previous iterations of the model. This was done using the correlation matrix. R's package, *corrplot*, was used to visualize the matrix. The correlation matrix plot shows the relationship between the predictor variables from the loan default dataset that were used in the final iteration of the logistic regression model.

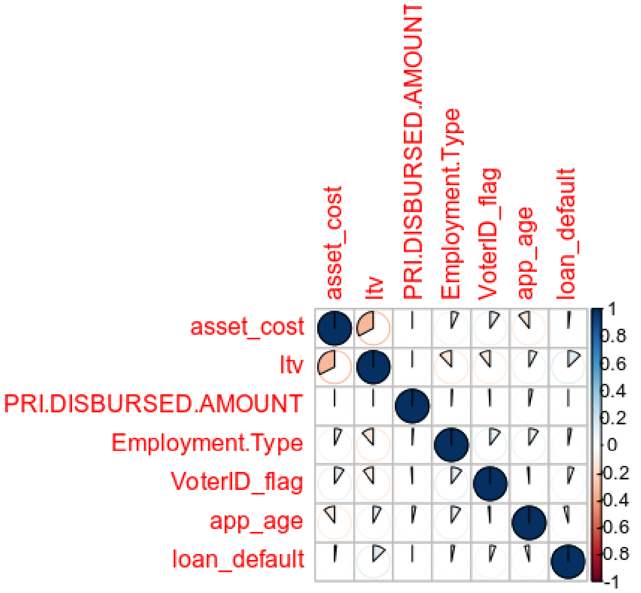


Fig. 2: Correlation Matrix Plot of the loan default prediction dataset.

4) Outliers: Outliers were removed, referring boxplots, from the datasets since they reduced the accuracy of the predictive models applied to them as you will see further in this report.

B. *Data Transformation*

To strive for more accurate predictions data transformation steps were carried out as detailed below.

1) Time and Dates: The 'time of day' predictor variable in the fire incident dispatch dataset was derived from the 'incident datetime' variable through binning, since it is intuitive that the response times might vary depending on the time of day. The plot below supports this claim and shows that it takes, on average, almost 20 seconds longer for the first unit to reach the scene of the incident than the time it would take for them to reach by night or midnight in the City of Manhattan. This could presumably be due to the increase in traffic during the morning.

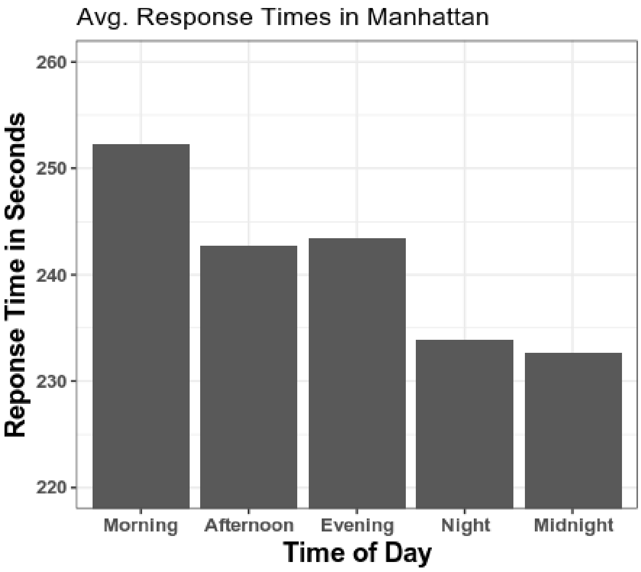


Fig. 3: Response times of fire services in Manhattan.

2) Feature Scaling: Feature scaling using the *scale()* function in R was done on all continuous predictor variables, like 'current balance' and 'disbursed amount', that were relatively high in magnitude when compared to the variables such as 'age' or 'active accounts'.

3) Under-Sampling: Under-sampling, using the *oven.sample()* function from the *ROSE* package in R, was performed on the loan default dataset. This was because the ratio of defaulters to non-defaulters was 1:3. The result of this sampling brought the ratio to an approximate 1:1 split. Multiclass under-sampling was attempted on the target variable in the 311 cases dataset using *RandUnderClassif()* from the *UBL* package from R, but this was not possible due to errors and no other reliable multiclass under-sampling function was found.

4) Train and Test Split: The resulting datasets were split into two subsets, one for training, and the other for testing the models. The *sample.split()* function from the *caTools* package from R was used to split each dataset as 75% for training and 25% for testing the models.

5) Dummy Coding: Since categorical variables from the fire incident dispatch dataset cannot be used as predictors in a Multiple Linear Regression model, they had to be dummy coded. This is a process of converting categories into dichotomous variables with 0's and 1's. This was accomplished using the *dummyVars()* function from R's *caret* package.

C. Data Mining Algorithms

The predictive models used on the datasets are as follows:

1) Naive Bayes Classification: (311 Cases)

The Naive Bayes Classification model is trained using the 311 cases dataset to predict whether a 311 case will be closed on the "Same day", "Within a Week", "Within a Month", or "Within a Year". It works on the principle of Bayes Theorem, and the ‘naive’ assumption is that the predictor variables are independent of each other. However, it performs just as well when this condition is not met, as it usually is so in the real world. An accuracy of 70% and a Kappa value of 0.45 was achieved. P-value of < 2.2e-16 was obtained.

2) KNN Classification: (311 Cases)

The KNN Classification model was trained on the 311 cases dataset to predict the resolution time class a 311 cases would fall in. It performs well with multiclass classification and was one of the reasons it was chosen for this prediction. An accuracy of 67.4% and a Kappa value of 0.26 was obtained. The p-value was 1.

3) Multiple Linear Regression: (Fire Incident Dispatch)

This model is trained on the data from the fire incident dispatch dataset to predict the response time (in seconds), which is the time from the alarm being rung in the fire station to the moment the first unit arrives on the scene of the incident. The R-Squared and Adjusted R-Squared values are both 15.9%. The p-value obtained was < 2.2e-16. The RMSE obtained before the removal of outliers was 158.21. The RMSE value obtained after was 134.28.

4) Random Forest Regression: (Fire Incident Dispatch)

The fire incident dispatch dataset was used to train the random forest regression model to predict the response time of the fire service department of New York. Before removing the outliers the RMSE was 154.1, and the RMSE value obtained after removing them was 133.23. It is noted that increasing the number of trees from 100 to 500 does not decrease the RMSE by much. The R-Squared value obtained is 0.16.

5) Logistic Regression: (Loan Default)

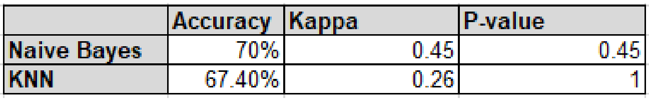
The loan default prediction dataset of vehicle loan applicants was used to train the logistic regression model to predict if an application would default their first loan repayment instalment. Before under-sampling, the accuracy of 78.3% with a Sensitivity of 1 and a Specificity of 0. After performing under-sampling using the *ROSE* package from R, the accuracy value obtained was 56.7%, but with an improved 0.63 as the Sensitivity and 0.51 as the Specificity.

II. Results and Discussions

The 5 models used are evaluated on a group of metrics. The interpretation of the results of the models is discussed below.

A. Dataset 1 (311 Cases):

Out of the two models used to predict the resolution time class for a 311 case, the Naïve Bayes model performs better than its counterpart, the KNN Classification model. The lower p-value, higher accuracy, and higher Kappa value of Naïve Bayes prove this. However, the p-value of Naïve Bayes is not significant and the prediction is more likely to occur by chance almost half the time. Hence, since the Naïve Bayes model was faster to train than KNN, the problems KNN posed with large number of classes in the categorical predictors, and its need for dummy coded variables, we can conclude that Naïve Bayes was better suited to predict the resolution time class of a 311 case.

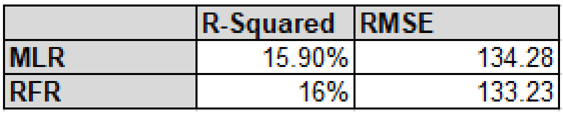


	Accuracy	Kappa	P-value
Naive Bayes	70%	0.45	0.45
KNN	67.40%	0.26	1

Fig. 4: Evaluation metrics of Naïve Bayes and KNN.

B. Dataset 2 (Fire Incident Dispatch):

Multiple Linear Regression (MLR) and the Random Forest Regression (RFR) were the two models trained using the NYFD fire incident dispatch data and used to predict the response times of the fire service department in having the first unit reach the incident scene. While the R-Squared and RMSE values for both the models are almost the same, their evaluation will have to include their degree of complexity. The time taken to run train the random forest model with 500 trees was substantially higher than the time taken to run the multiple linear regression model. However, the requirement for categorical predictor variables to be dummy coded in a multiple linear regression becomes incredibly complex to interpret the model coefficients, especially if the dataset in question has many categorical variables, each with many categories. In the end, it comes down to the speed vs interpretability trade-off. Also considering the root of the MSE is very high, it indicates the poor predictive performance of both the models.



	R-Squared	RMSE
MLR	15.90%	134.28
RFR	16%	133.23

Fig. 5: Evaluation metrics of multiple linear regression and random forest regression.

The plot below indicates the poor performance of the MLR model in which it is seen that response times less than a minute are predicted to be 150-350 seconds. This is likely due to the high values in the dataset that skew the distribution. Hence, from the plot, it is clear that the model is incapable of predicting extreme response times at either end of the spectrum.

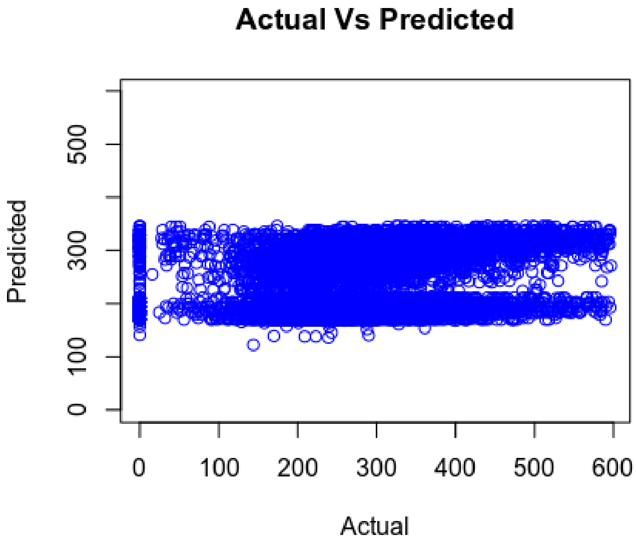
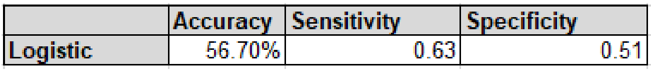


Fig. 6: Actual vs predicted plot of multiple linear regression.

C. Dataset 3 (Loan Default Prediction):

The loan default prediction dataset from L&T has been used to train the logistic regression model which results in the metric values shown in the table below. However, in the context of predicting loan defaulters, this model is poor in its prediction. With an accuracy of 56.7%, almost half the time the predictions would be inaccurate. And with low sensitivity and specificity values, this model would not be of much help in bringing about a considerable improvement to the loaner by predicting loan defaulters.



	Accuracy	Sensitivity	Specificity
Logistic	56.70%	0.63	0.51

Fig. : Evaluation metrics of logistic regression.

IV. CONCLUSION AND FUTURE WORK

In this paper, a total of 5 models were trained with a total of 3 different datasets from different domains. The objective to train the 1st and 2nd model with the 1st dataset, the 3rd and 4th with the 2nd dataset, and the 5th model with the 3rd dataset. In the multiclass classification group of methods, we see that KNN. Prior to deciding on using the KNN model, a decision tree and SVM were trialed. Since there are many classes in the predictor categorical variables, this becomes computationally expensive in decision trees and SVM. Due to the lack of computational and memory resources on the local system, the attempt to use these methods to predict the resolution time of the 311 cases was aborted. KNN could also not run for a high value of k or would usually end up getting a tie, which happens when there is more than one class having the joint highest number of neighbours to the observation to be predicted.

There were some limitations associated with the performance of the models. The plan for the future would be to improve the accuracy of the models by using more efficient feature selection methods to train the models rather than relying only on correlation coefficients or p-values of the features used. A reliable R package that can perform multiclass under-sampling will be used to improve the training dataset. Boosting algorithms such as XGBoost will be researched and used to improve model performances.

V. REFERENCES

[1] J. Goderie, B. M. Georgsson, B. v. Graafeiland and A. Bacchelli, "ETA: Estimated Time of Answer Predicting Response Time in Stack Overflow," 2015 IEEE/ACM 12th Working Conference on Mining Software Repositories, Florence, 2015, pp. 414-417.

[2] S. Boyles, D. Fajardo, S. Waller and A. Professor, “Naive Bayesian classifier for incident duration prediction”, 2007.

[3] Y. Lin, Y. Zhang, I. Lin and C. Chang, "Predicting logistics delivery demand with deep neural networks," *2018 7th International Conference on Industrial Technology and Management (ICITM)*, Oxford, 2018, pp. 294-297.

[4] W. Zhou, L. Tang, C. Zeng, T. Li, L. Shwartz and G. Ya. Grabarnik, "Resolution Recommendation for Event Tickets in Service Management," in *IEEE Transactions on Network and Service Management*, vol. 13, no. 4, pp. 954-967, Dec. 2016.

[5] H. Xu, H. Zhang, F. Zong and W. Qi, "Traffic incident Duration analysis using Hazard duration model," *Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)*, Changchun, 2011, pp. 1301-1304.

[6] R. Eshleman and H. Yang, ""Hey #311, Come Clean My Street!": A Spatio-temporal Sentiment Analysis of Twitter Data and 311 Civil Complaints," *2014 IEEE Fourth International Conference on Big Data and Cloud Computing*, Sydney, NSW, 2014, pp. 477-484.

[7] L. Ren, H. Chang and Y. Yi, "An Improved Binary Tree SVM Classification Algorithm Based on Bayesian," *2009 Asia-Pacific Conference on Information Processing*, Shenzhen, 2009, pp. 178-181.

[8] S. Singh and U. Soni, "Predicting Order Lead Time for Just in Time production system using various Machine Learning Algorithms: A Case Study," *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2019, pp. 422-425.

[9] S. Deshpande, K. Dheenadayalan, G. Srinivasaraghavan and V. Muralidhara, "Filer Response Time Prediction Using Adaptively-Learned Forecasting Models Based on Counter Time Series Data," *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, CA, 2016, pp. 13-18.

[10] J. J. Rebollo and H. Balakrishnan, “Characterization and Prediction of Air Traffic Delays.” *Transportation Research Part C: Emerging Technologies 44*, 2014, pp. 231–241.

[11] S. Handley, P. Langley and F. A. Rauscher, “Learning to Predict the Duration of an Automobile Trip”, 1998.

[12] C. Huang and L. Ku, "EmotionPush: Emotion and Response Time Prediction Towards Human-Like Chatbots," *2018 IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, United Arab Emirates, 2018, pp. 206-212.

[13] Y. He, S. Blandin, L. Wynter and B. Trager, "Analysis and Real-Time Prediction of Local Incident Impact on

Transportation Networks," *2014 IEEE International Conference on Data Mining Workshop*, Shenzhen, 2014, pp. 158-166.

[14] B. Van de Vyvere, K. D'Haene, K. D'Haene, P. Colpaert, and R. Verborgh, "Predicting phase durations of traffic lights using live Open Traffic Lights data," in *Proceedings of the First International Workshop on Semantics for Transport*, 2019.

[15] R. Kumar, S. M. Naik, V. D. Naik, S. Shiralli, Sunil V.G and M. Husain, "Predicting clicks: CTR estimation of advertisements using Logistic Regression classifier," *2015 IEEE International Advance Computing Conference (IACC)*, Banglore, 2015, pp. 1134-1138.

[16] E. Afatmirni et al., "Predicting refibrillation from pre-shock waveforms in optimizing cardiac resuscitation," *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Boston, MA, 2011, pp. 251-254.

[17] Y. Sayjadah, I. A. T. Hashem, F. Alotaibi and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques," *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, Subang Jaya, Malaysia, 2018, pp. 1-4.

[18] S. H. Ebenuwa, M. S. Sharif, M. Alazab and A. Al-Nemrat, "Variance Ranking Attributes Selection Techniques for Binary Classification Problem in Imbalance Data," in *IEEE Access*, vol. 7, pp. 24649-24666, 2019.

[19] T. S. Brisimi, T. Xu, T. Wang, W. Dai, W. G. Adams and I. C. Paschalidis, "Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach," in *Proceedings of the IEEE*, vol. 106, no. 4, pp. 690-707, April 2018.

[20] G. Sudhamathy and C. J. Venkateswaran, "Analytics using R for predicting credit defaulters," *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, Coimbatore, 2016, pp. 66-71.

VI. Footnotes

[1] Junior, Vic Lang'at, "The 10 Largest Cities in India." WorldAtlas. [Online]. Available: https://www.worldatlas.com/articles/the-10-largest-cities-in-india.html [Accessed on: Dec. 15, 2019].

[2] R. McNamara, "Cork ambulances failing to meet response-time targets," EchoLive.ie: Feb. 6, 2018. [Online]. Available: https://www.echolive.ie/corknews/Cork-ambulances-failing-to-meet-response-time-targets-fced5586-a468-4736-8181-a2c8b8fc90e7-ds [Accessed on: Nov. 8, 2019].

[3] J. Brennan, "AIB chief says loan defaulters add to costs for all customers," The Irish Times: Jun. 20, 2018. [Online]. Available: https://www.irishtimes.com/business/financial-services/aib-chief-says-loan-defaulters-add-to-costs-for-all-customers-1.3537702 [Accessed on: Dec. 15, 2019].

[4] AndyM (San Francisco 311), "311 Cases" Nov 8, 2019. [Online]. Available: https://data.sfgov.org/City-Infrastructure/311-Cases/vw6y-z8j6 [Accessed on: Dec. 15, 2019].

[5] Fire Department of New York City (FDNY), "Fire Incident Dispatch Data," Mar. 21, 2019. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/Fire-Incident-Dispatch-Data/8m42-w767 [Accessed on: Dec. 15, 2019].

[6] M. Dhaker, "L&T Vehicle Loan Default Prediction,". [Online]. Available: https://www.kaggle.com/mamtadhaker/lt-vehicle-loan-default-prediction [Accessed on: Dec. 15, 2019].

[7] Anderson Roges et al., "Classification of Power Quality Considering Voltage Sags in Distribution Systems Using KDD Process," May 2015. [Online]. Available: http://www.scielo.br/scielo.php?pid=S0101-74382015000200329&script=sci_arttext [Accessed on: Dec. 15, 2019].