Classification of Credit Card Defaulters, Influence on Air pollutants and Hotel cancellation prediction using Machine Learning Algorithms

Victory Chimamaka Uwaoma
MSc. Data Analytic
National College of Ireland
Dublin, Ireland
x19210931

Abstract— **This project tackles the probability of a credit card default by clients using logistics regression and random forest algorithm. The end result shows the random forest with an accuracy of 82% and a higher area under the curve of 0.765 is better at predicting clients who were likely to default with their credit card. The second objective is to determine the relationship between the levels of air pollutants and the weather conditions. This was achieved by modeling the different air pollutants by the weather conditions using multiple linear regression. The end result among the pollutant PM2.5, NO2 and CO showed that the NO2 has a relationship with the weather conditions with the least root mean squared error (RMSE) of 45.9800646. The third objective is to determine the probability of a client canceling a hotel booking previously made using k-nearest neighbor (kNN) and support vector machine (SVM). Again, the end result of this analysis shows the kNN as a better predictor with an accuracy of 80%.**

Keywords— Algorithms, Multiple linear regression, Logistics regression, KNN, Random forest, Accuracy, SVM

I. INTRODUCTION

Gone are the days when programmers would tell a computer how to solve a problem. We are in the machine learning era where computers are left to solve problems on their own by finding patterns for solutions. The study of the unseen trends and patterns makes it easier to predict and prevent future problems from happening. Therefore, the need for machine learning is crucial especially in the areas where large data (big data) is involved. This project tackles areas where you can apply data mining and machine learning algorithms to produce maximum result.

A. Predicting Credit card default

In today's world, everyone is allowed to own a credit card. The ability to pay for a service online using a credit card is a privilege that all credit card holders enjoy, but problem exists when those who own a credit card spend to the maximum without been able to refund the money back to the credit card company. Hence a credit card system is formed in other to know the credit details of client and know when to award credit scores to deserving clients. This record also holds the credit history of the clients and can be used to know when a client is liable to default on his credit card. The traditional way wherein banks generally approved or rejected a customer's credit was to use presumptuous techniques in which some predetermined factors are taken into consideration [1], but with the help of a predictive systems from machine learning, this will make it fast, easy and efficient to predict a credit card holder as defaulters or not. Knowing whether or not a customer defaults on his credit card payment is a yes or no situation which means that it has only two outcomes. Therefore, it will be suitable for a classification-based algorithm. In this project, the logistics regression and the random forest will be used in determining if a client will default on his credit card or not.

B. Influence on weather on Air Pollution

Pollution is one of the major causes in our environment today. Most of these pollutions are caused by the gases present in the air which is as a result of the modern industries and refineries in our society. People have been paying more and more attention to air pollution in recent years since it directly impacts the health and everyday life of individuals. This affects the health of individuals causing health related problems such as lung cancer, heart related problems and even death in human, plants, and animals. Efficient prediction of air pollution has become one of the issues faced by researchers [2]. The aim of this research project is to determine the level of air pollutant using the weather conditions such as temperature, pressure, dew point temperature, rain and windspeed. Machine learning will be suitable for this regression problem analysis. The multiple linear regression will be used in modeling the air pollutants.

C. Predicting Hotel cancellation

Several considerations are taking in when booking a hotel, especially if it is a place far away from home. In order to avoid disappointment and inconveniences, it is better to know the best time to book a hotel for reservation. Thus, having the right information about booking a hotel will be of great comfort especially when one moves from place to place. But on the other hand, in other for the hotel management system to be able to predict if a client will possibly cancel on their reservation made is important as it has a negative effect on their revenue [3]. The outcome of a cancellation is either a cancel or not cancel, therefore, classification-based algorithm such as the k nearest neighbor (kNN) and the support vector machine will be used to determine the outcome.

II. RELATED WORKS

Machine learning algorithms have been in existence for quite a while and have been used to give understanding on data in our day-to-day lives in diverse areas such as forecasting, medical diagnosis, stock market and more to make critical decisions and predictions. Different machine learning algorithm has now and then led to similar and sometimes distinct insights as some machine learning

algorithms perform a lot better depending on the data on which the algorithm is been carried out.
Several related works have been done in the area of credit default, pollution, and hotel prediction.

A. Predicting Credit card default

Machine learning methods such as random forest have been used to solve issues related to credit cards, for example, [4] carried out a detection of credit card fraud which was centered on fraudulent transactions. The random forest had an accuracy of 90%. In an effort to determine the credit risk in the banking industry, [5] carried out analysis and the end result showed the random forest with an accuracy of 82% and an area under the curve value of 77% as the best classifier. Another analysis was carried out by [6] in the same area to prevent the inability of customers to pay their card balances monthly. Data mining-based failure prevention system was done to avoid the risk. The random forest again was used in this analysis. The model produced an F-measure of 0.89 with an area under the curve of 0.947 and a predictive accuracy of 89.01%. The random forest algorithm has also been used in other areas. [7] used the random forest in a study to identify the source of disturbance from a defined list which has adverse effects on events of power quality. The random forest had the highest accuracy of 98.68%.
Logistics regression has been used in predicting credit card defaulters. [1] made use of three prediction models to aid in credit validity assessment during the credit application process. The logistics regression model was seen to produce an accuracy of 74.56% which makes it useful in predicting. Again, [8] used clustering-launched classification (CLC) and support vector machines in predicting credit scoring on two datasets. The CLC outperformed all the other models with better predictive accuracies on both datasets. Lastly, [9] used logistic regression model to predict credit card consumer churning based on a cluster-stratified parameters appropriate for data set imbalance, it had a predictive accuracy of 73.5%.

B. Influence on weather on Air Pollution

In the area of pollution, machine learning algorithms have been used in this study. The Multiple linear regression machine learning technique involves the study of a target variable in relations to two or more predictor variables. A research done in Malaysia by [10] shows that multiple linear regression has been used to predict pollution such as particulate matter (PM10). The analysis shows that MLR has a predictive accuracy of 94% which is suitable for predicting air pollution. [11] used multiple linear regression based on principle component to predict the ozone concentration in Northern Portugal. The result showed that the addition of principle components as input improved the model by removing the data collinearity and reducing the complexity giving it a predictive accuracy of 70%. [12] used regression models to predict traffic-related air pollution in Rome. Using emissions and land-use data, the result shows that the emissions data do not really improve the model while the land-use regression model accounts for the air pollution levels related to traffic with reasonable accuracy of 69%. Multiple linear regression has been used in several study to tackle different areas. For example, the multiple linear regression method was also used to predict the Pb2 + ion and the Cd2 + ion using two ionic selective electrodes as predictors. The multiple regression model was found to have accurately displayed an R squared value of 0.990020 [13]. The multiple regression technique was used to research the quantification of bone-setting manipulation of information in Chinese medicine [14].
Another example where the multiple linear has been used was to predict and forecast demand for bicycles in Washington DC. Multiple linear regression provided a weak turn-out in terms of its predictive accuracy, but there was a decent relationship between the factors considered, this was due to one of the predictor variables named weather having a lot of dummy variables [15]. Other machine learning algorithms have been used in predicting air pollution. [16] proposed a study to find an alternative way to track and characterize air quality using integrated gas sensors and the development of predictive models using machine learning algorithms. This was done by developing a model for the optimized sensors using the temperature and relative humidity sensors DHT 11, MQ2, MQ5 and MQ135. Using five models, k-nearest neighbors (KNN), support vector machine (SVM), Naïve-Bayesian classifier, random forest, and neural network. The result showed the neutral network with 99.56%.

C. Predicting Hotel cancellation

In the area of predicting hotel cancellation, similar research has been carried out. Nuno Antonio et al [3] carried out a research using different machine learning algorithms on predicting hotel booking cancellations to decrease uncertainty and increase revenue. This research was done to predict why booking cancellations occurred and how to prevent it in order to increase the revenue management. The analysis was carried out using the CRISP-DM methodology and was done on two different hotels. The result of the analysis had an accuracy above 0.84. Again [17] proposed two models, Back propagation neural network (BPN), general regression neural network (GRNN) in predicting hotel cancellation by customers. The research reports showed that both models have significant predictive capabilities of 80% and 87.14% specificity respectively which can help in the system. In another research, [18] created models to detect changes in patterns of cancelations over time and learn from the hits and inaccuracies predicted in previous days and assessing the effectiveness of predictions over time. The tests surpassed 84% in accuracy, 82% in precision and 88% in Area Under the Curve (AUC). In [19] the study showed cancelation policies and their role in shaping the deal-seeking behavior of travelers, explored the effect of cancelation fees and deadlines using non-parametric multiple pairwise comparisons, and multinomial logit regression models. The results suggest that the cancelation deadline influenced the actions of the participants, although there was no statistically significant effect on the size of the cancelation charge. Using data mining-based cancellation rate forecasting models [20] tried to find the probability why bookings may be cancelled or why booked customer does not turn up at the time of service. The study showed that the different models where better at different time point. Lastly in India, a similar research was carried out, but it was done in the area of flight cancellation prediction [21]. The analysis was done using classification-based models. The end result showed that the random forest has the best precision of 100%.

III. METHODOLOGY

This project will adopt the Knowledge Discovery in Databases (KDD) methodology and this will be applied to the three separate analyses that is conducted in this project. The figure below shows an outline of the steps of how the KDD process works.

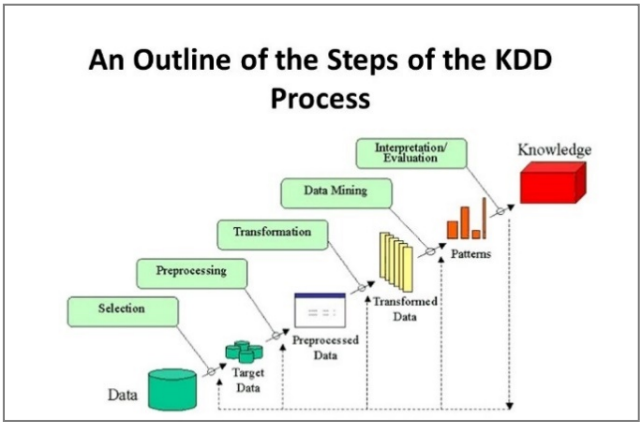

Figure 1: KDD diagrammatic representation

First the data to be analyzed will be sourced and cleaned, after proper cleaning of the data, only data points relevant to the analysis are extracted and data mining methods such as regression, classification, and clustering are then applied to the data to detect patterns and lastly the outcomes of the techniques applied are then analyzed in order to determine their performance on the data.
The process according to the KDD methodology is defined as follows:

- Data sourcing
- Data preprocessing and transformation
- Data selection
- Implementation of Algorithms
- Result

A. Credit card default

The project's first analysis involves the classification of credit card default by clients. Using the logistics regression and random forest model on the credit card default dataset, analysis will be done in order to determine the patterns. The process will be done using the KDD process.

Data sourcing
The credit card default dataset was gotten from the UCI Machine learning repository: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. The dataset contains the credit card information of default payments of clients in Taiwan as well as other attributes. It is comprised of 25 attributes and 30000 instances. The following table gives more details on the attributes and their meanings

Table 1: Credit card default

Variable	Type	Description
ID	Numeric	Client ID
X1	Integer	Amount of the given credit (NT dollar)
X2	Categorical	Gender (1 = male; 2 = female)
X3	Categorical	Education (1= graduate school, 2 = university, 3 = high school, 4 = others)
X4	Categorical	Marital status (1 = married, 2= single, 3= others)
X5	Integer	Age (Year)
X6	Categorical	Repayment status in September 2005
X7	Categorical	Repayment status in August 2005
X8	Categorical	Repayment status in July 2005
X9	Categorical	Repayment status in June 2005
X10	Categorical	Repayment status in May 2005
X11	Categorical	Repayment status in April 2005
X12	Integer	Amount of bill statement in September 2005
X13	Integer	Amount of bill statement in August 2005
X14	Integer	Amount of bill statement in July 2005
X15	Integer	Amount of bill statement in June 2005
X16	Integer	Amount of bill statement in May 2005
X17	Integer	Amount of bill statement in April 2005
X18	Integer	Amount paid in September 2005
X19	Integer	Amount paid in August 2005
X20	Integer	Amount paid in July 2005
X21	Integer	Amount paid in June 2005
X22	Integer	Amount paid in May 2005
X23	Integer	Amount paid in April 2005
Y	Categorical	Payment default (1= yes, 0 = no)

Data preprocessing and transformation
In this phase, the sourced data was read into the RStudio software and was cleaned and transformed as follows. Firstly, the dataset had a first column which was not of importance, it was programmatically removed. Then the target variable had space in its name which was also changed to an appropriate naming conversion. Lastly, using the sapply function, all the variables except the target variable where changed to characters and then to numeric. This was done because the algorithms which was to be used could only work with numerical values. The target variable was left as a factor because it contained categorical classes. The data was checked for NA values, which none was discovered hence no further preprocessing was done.

Data Selection
Only the variables relevant for the analysis were selected in this stage. The selection was done using a heatmap. The heatmap was used to visualize the relationships between target variable and its predictor variables.

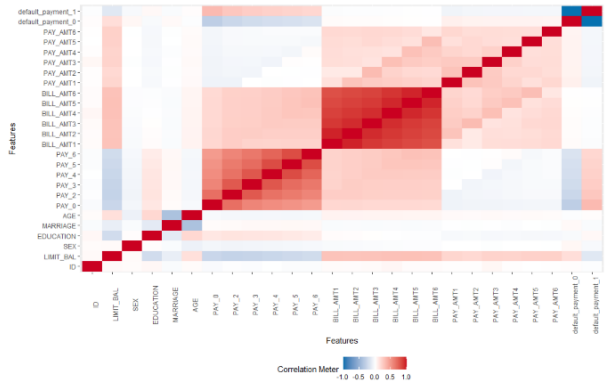

Figure 2: Heatmap

Close examination at the heatmap above in figure 2, shows weak correlations of 'ID', 'AGE', 'BILL_AMT2', 'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', and 'BILL_AMT6' to the target variable, therefore, they were all removed from the analysis for optimal result.

Implementation of Algorithms

After loading the appropriate libraries, the dimension and structure of the data was checked to make sure the data was properly entered. The logistics and random forest algorithms where used in fitting the credit card default data. The data was split into 70% for training and 30% for testing for the analysis. Logistics classifier was first used in training the dataset, after training the model, it was then tested using the testing dataset for predictions. Below shows the summary of the logistics model.

Deviance Residuals				
Min	1Q	Median	3Q	Max
-3.1006	-0.7004	-0.5485	-0.2969	3.7382

Since the target variable was in categorical format, a decision rule was set, that if the data was greater than 0.5 it should be predicted as 1 (default) else 0 (not default). Performance evaluation such as confusion matrix and roc curve.

The confusion matrix below was used to show the true positive and false positive values

Reference		
Prediction	0	1
0	6838	1540
1	168	454

The table shows 6838 cases did not default with their credit card and were predicted correctly (true positive), 1540 defaulted but was predicted wrongly as not defaulting (false positive), while 168 who did not default were predicted they did (false negative) and lastly, 454 cases who defaulted where predicted correctly (true negative). The confusion matrix also calculated the accuracy of the model as 0.8102 which is roughly 81% as well as the Kappa value 0.2702 which is used in measuring the inter-rater reliability of the model, the Kappa is considered as fair measure. The sensitivity and specificity of the model were 0.9760 and 0.2277 respectively. The receiver operating characteristic (ROC) curve was used as another evaluation measure for the model. This helps to show the graphical representation between the sensitivity and specificity. Below is the ROC curve for the logistics classifier.

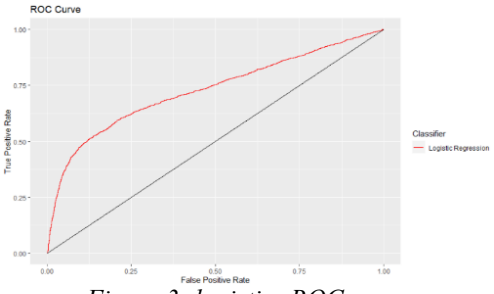

Figure 3: logistics ROC curve

The second classifier used in fitting the model was the random forest classifier. It was used on the same testing and training dimensions as the logistics regression. After the model has been trained and tested, a plot was done on the random forest using the trained data. The figure 4 below shows the plot.

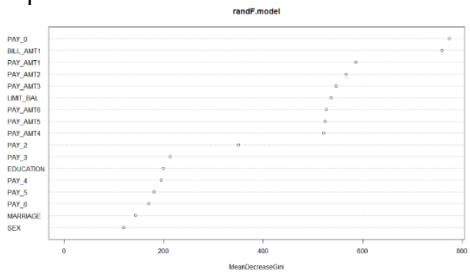

Figure 4: A plot showing the random forest

The confusion matrix was used to show prediction of the random forest model.

Reference		
Prediction	0	1
0	6648	358
1	1241	753

This shows that 6648 cases where predicted correctly as not defaulting, 358 cases who defaulted where predicted as non-default, 1241 cases who did not default where predicted as default and lastly, 753 cases who defaulted were predicted correctly. 0.8223 was calculated as the accuracy which is approximately 82%. The kappa value was calculated to be 0.388 which is a fair measure, the sensitivity and specificity were shown to be 0.8427 and 0.6778 respectively. The figure 5 below shows the ROC curve of the random forest model

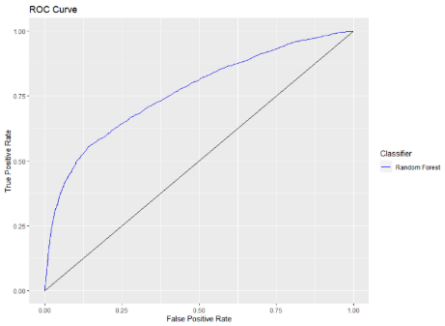

Figure 5: ROC curve of random forest model

Results

The aim of this analysis was to classify the credit card into defaulters and non-defaulters using logistics and random forest classifiers and determine which algorithm performed better. A combination of the logistics regression and random forest ROC was done for a better result reference.

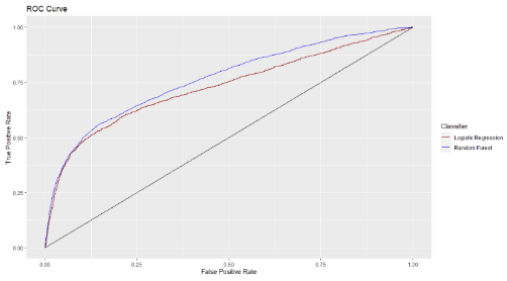

Figure 6: ROC curve of logistics and random forest model

From figure 6 above, the random forest classifier shows to have a greater Area Under the Curve (AUC) with 0.765 compared to the logistics regression which has 0.728, this shows the random forest classifier as a better predictor for the classification of credit card default.

B. Beijing Air-Quality Data

The second analysis talks about relationship of the weather conditions on air pollutants. Using the multiple linear regression, the analysis will be done in order to determine the patterns. The process will be done using the KDD process.

Data sourcing

The Beijing Air-quality dataset was gotten from the UCI Machine learning repository: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data. The dataset was collected from 12 nationally regulated air quality monitoring sites on hourly air pollutants. The statistics on the air quality comes from the Beijing Municipal Environmental Monitoring Centre and this data comprises of 18 attributes and 420768 instances. But this analysis will make use of only one station (Tiantan) which has a sample size 35064. The table gives more information on the data.

Table 2: Pollution table

Variable	Type	Description
No	Integer	Row number
Month	String	Month of the data in the row
Day	Integer	Day of data in the row
Hour	Integer	Hour of data in the row
PM2.5	Integer	PM2.5 concentration (ug/m^3)
PM10	Integer	PM10 concentration (ug/m^3)
SO2	Integer	SO2 concentration (ug/m^3)
NO2	Integer	NO2 concentration (ug/m^3)
CO	Integer	CO concentration (ug/m^3)
O3	Integer	O3 concentration (ug/m^3)
TEMP	Continuous	Temperature in degree Celsius
PRES	Continuous	Pressure (hPa)
DEWP	Continuous	Dew point temperature (degree Celsius)
RAIN	Integer	Precipitation (mm)
wd	String	Wind Direction
WSPM	Continuous	Wind speed (m/s)
Station	String	Name of air quality monitoring site

Data preprocessing and transformation

The sourced data was cleaned and transformed as follows. First, the dimension and structure of the data was checked. Using the sapply and is.na function, the data was checked for NA values of which some were discovered. The NAs found were programmatically removed because it did not amount to 3% of the data.

Data Selection

In this phase, variables suitable for the analysis were selected. The selection was done by checking their correlation using the cor and pairs.panels function was used to show relationships between dependent and independent variables.

	PM2.5	NO2	CO	TEMP	PRES	DEWP	RAIN	WSPM
PM2.5	1.0000000000	0.66772300	0.79912119	-0.14609609	-0.0004972861	0.11932169	-0.01529090	-0.29743821
NO2	0.6677230029	1.00000000	0.71663502	-0.32120563	0.1815242342	-0.07824441	-0.04844370	-0.41396429
CO	0.7991211886	0.71663502	1.00000000	-0.31666640	0.1464027044	-0.03254889	-0.01461271	-0.32930357
TEMP	-0.1460960914	-0.32120563	-0.31666640	1.00000000	-0.8337042490	0.82148858	0.03885590	0.03839147
PRES	-0.0004972861	0.18152423	0.14640270	-0.83370425	1.0000000000	-0.77150745	-0.06727670	0.05082284
DEWP	0.1193216900	-0.07824441	-0.03254889	0.82148858	-0.7715074520	1.00000000	0.08889869	-0.28631347
RAIN	-0.0152909020	-0.04844370	-0.01461271	0.03885590	-0.0672766990	0.08889869	1.00000000	0.02587387
WSPM	-0.2974382075	-0.41396429	-0.32930357	0.03839147	0.0508228431	-0.28631347	0.02587387	1.00000000

Figure 7: Correlation between variables (using cor function)

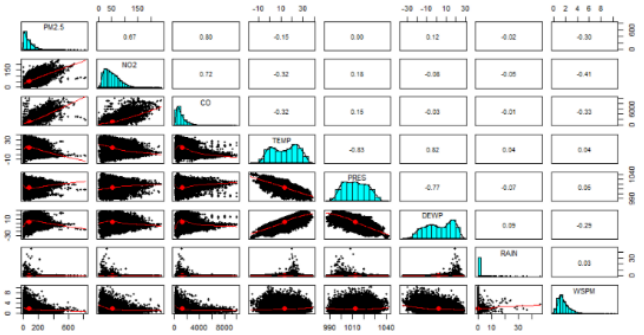

Figure 8: Correlation between variables (using pairs.panels function)

The analysis was modeled using three different air pollutants such as PM2.5, NO2 and CO as the target variables and the weather conditions as the predictors. This analysis was done using the multiple linear regression model on all three models. The first model has the PM2.5 as the target variable and the weather conditions as the predictor variables (temp, press, dewp, rain and wspm). The second model has the NO2

as its target variable and all others as the predictor variables and lastly, CO was the target variable for the third model.

Implementation of Algorithms

The appropriate libraries were loaded, the data was read in and the dimension and structured of the data was checked to make sure the data was properly entered. The multiple linear classifier was used on the three models. The data was split into 70% for training and 30% for testing. The multiple linear was used in training the dataset. After training with all the models, they were then tested using the testing dataset for predictions. The summary of all the multiple linear model are shown below.

For the first model (PM2.5)

The PM2.5 pollutant was used to fit the weather conditions using the multiple linear regression model the summary is shown below:

Residuals				
Min	1Q	Median	3Q	Max
-142.43	-46.68	-12.93	29.14	691.51

Coefficients				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1945.74564	87.92185	22.130	<2e-16 ***
TEMP	-5.78176	0.09543	-60.588	<2e-16 ***
PRES	-1.76223	0.08614	-20.458	<2e-16 ***
DEWP	3.50908	0.07367	47.632	<2e-16 ***
RAIN	-4.75173	0.57190	-8.309	<2e-16 ***
WSPM	-5.02427	0.44152	-11.380	<2e-16 ***

Confidence Intervals	2.5 %	97.5 %
Intercept	1773.412905	21188.078374
TEMP	-5.968802	-5.594715
PRES	-1.931072	-1.593398
DEWP	3.364684	3.653483
RAIN	-5.872697	-3.630763
WSPM	-5.889670	-4.158863

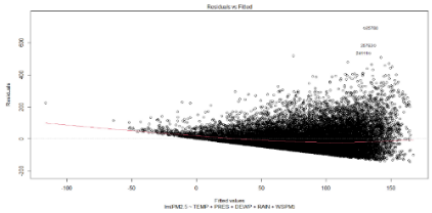

Figure 9 residual vs fitted

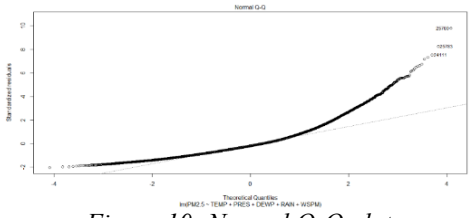

Figure 10: Normal Q-Q plot

Figure 9 and figure 10 above shows the residual vs the fitted plot and the normal q-q plot for the model.

The summary after the testing data was done on the model is given below:

	Residuals				
Min	1Q	Median	Mean	3Q	Max
-59.35	54.92	85.93	81.40	111.23	170.75

RMSE	R squared
70.7973553	0.2184179

The multiple linear regression for the first model (PM2.5) produced an R square value of 0.218 which amounts to roughly 21% accuracy and a root mean squared error of 70.7973553

For the second model (NO2)

The NO2 pollutant was used to fit the weather conditions using the multiple linear regression model the summary is shown below:

Residuals				
Min	1Q	Median	3Q	Max
-64.276	-18.170	-4.732	13.321	186.519

Coefficients				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	546.69891	33.67206	16.236	<2e-16 ***
TEMP	-1.71713	0.03655	-46.985	<2e-16 ***
PRES	-0.45107	0.03299	-13.673	<2e-16 ***
DEWP	0.52959	0.02821	18.770	<2e-16 ***
RAIN	-1.76557	0.21903	-8.061	7.93e-16 ***
WSPM	-7.87790	0.16909	-46.590	<2e-16 ***

The model has a confidence interval as shown below

Confidence Intervals	2.5 %	97.5 %
Intercept	480.6994070	612.6984098
TEMP	-1.7887664	-1.6454995
PRES	-0.5157290	-0.3864075
DEWP	0.4742905	0.5848939
RAIN	-2.1948760	-1.3362665
WSPM	-8.2093265	-7.5464673

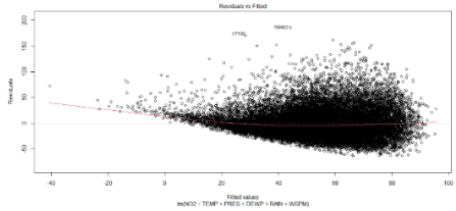

Figure 11: residual vs fitted

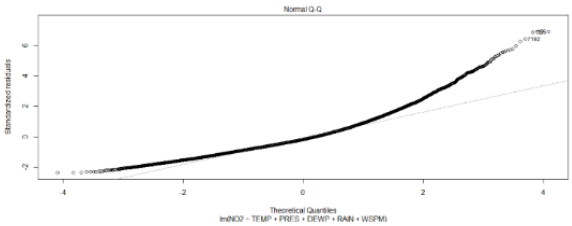

Figure 12: Normal Q-Q plot

Figure 11 and figure 12 above shows the residual vs the fitted plot and the normal q-q plot for the model.

The summary after the testing data was done on the model is given below:

	Residuals				
Min	1Q	Median	Mean	3Q	Max
-17.17	41.70	54.37	53.08	65.40	94.45

RMSE	R squared
45.9800646	0.2184179

The multiple linear regression for the second model (NO2) produced an R square value of 0.218 which amounts to roughly 21% accuracy and a root mean squared error of 45.9800646

For the third model (CO)

The CO pollutant was used to fit the weather conditions using the multiple linear regression model the summary is shown below:

Residuals				
Min	1Q	Median	3Q	Max
-2236.1	-601.9	-146.5	365.3	8732.7

Coefficients				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27542.811	1232.308	22.35	<2e-16 ***
TEMP	-93.304	1.338	-69.76	<2e-16 ***
PRES	-24.538	1.207	-20.32	<2e-16 ***
DEWP	43.442	1.033	42.07	<2e-16 ***
RAIN	-50.180	8.016	-6.26	3.91e-16 ***
WSPM	-124.305	6.188	-20.09	<2e-16 ***

The model has a confidence interval as shown below

Confidence Intervals	2.5 %	97.5 %
Intercept	25127.40349	29958.21783
TEMP	-95.92549	-90.68230
PRES	-26.90420	-22.17138
DEWP	41.41807	45.46585
RAIN	-65.89150	-34.46867
WSPM	-136.43473	-112.17583

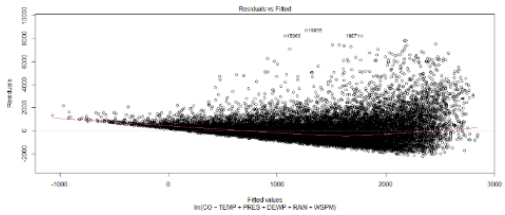

Figure 13: Residual vs fitted

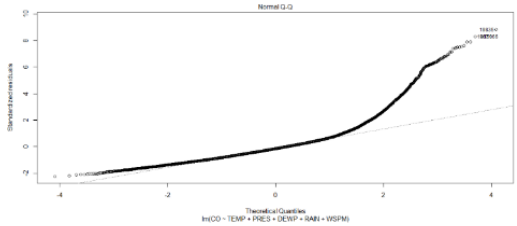

Figure 14: Normal Q-Q plot

Figure 13 and figure 14 above shows the residual vs the fitted plot and the normal q-q plot for the model.

The summary after the testing data was done on the model is given below:

	Residuals				
Min	1Q	Median	Mean	3Q	Max
-896.8	863.0	1332.7	1294.8	1740.3	2844.1

RMSE	R squared
70.7973553	0.2184179

The multiple linear regression for the third model (CO) produced an R square value of 0.218 which amounts to roughly 21% accuracy and a root mean squared error of 70.7973553

Results

The aim of this analysis was to check the relationship between the weather conditions on the levels of air pollutant using multiple linear regression and determine which pollutant has the no influence. The end result of the analysis showed that all three models had an R square value of 0.2184179 which accounts for 21% of the accuracy. While PM2.5 had a root squared mean error of 70.79 and NO2 a 45.98 value, the CO pollutant had an RMSE of 1652.97. Therefore, NO2 has the most influence on the weather conditions since it has the lowest root squared mean of 45.98.

C. Hotel Booking

The third analysis involves predicting the possibility of a client canceling a hotel reservation or not. This analysis will be done using k nearest neighbor and support vector machine. Again, the KDD methodology will be used.

Data sourcing

The hotel booking dataset was gotten from Kaggle repository: https://www.kaggle.com/jessemostipak/hotel-booking-demand/data#. The dataset provides details about possibility of cancelling for a city hotel and a resort hotel booked and contains information such as when the reservation was. It is comprised of 32 attributes and 119390 rows. The following table gives more details on the attributes and their meanings

Table 3: Hotel booking

Variable	Type	Description
Hotel	Categorical	Types of hotel
ADR	Numeric	Average Daily Rate
Adults	Integer	Number of adults
Agents	Categorical	ID of the travel agency that made the booking
ArrivalDateDayOfMonth	Integer	Day of the month of the arrival date
ArrivalDateMonth	Categorical	Month of arrival date with 12 categories: "Jan" to "Dec"
ArrivalDateWeekNumber	Integer	Week number of the arrival date
ArrivalDateYear	Integer	Year of arrival date
AssignedRoomType	Categorical	Code for the type of room assigned to the booking
Babies	Integer	Number of babies
BookingChanges	Integer	Number of changes
Children	Integer	Number of children
Company		ID of the company/entity that made the booking
Country	Categorical	Country of origin.
CustomerType	Categorical	Type of booking
DaysInWaitingList	Integer	Number of days the booking was in the waiting list before it was confirmed to the customer
DepositType	Categorical	Indication on if the customer made a deposit to guarantee the booking.
DistributionChannel	Categorical	Booking distribution channel.
IsCanceled	Categorical	Value indicating if the booking was canceled (1) or not (0)
IsRepeatedGuest	Categorical	Value indicating if the booking name was from a repeated guest (1) or not (0)
LeadTime	Integer	Number of days that elapsed
MarketSegment	Categorical	Market segment designation
Meal	Categorical	Type of meal booked
PreviousBookingsNotCanceled	Integer	Number of previous bookings not cancelled by the customer prior to the current booking
PreviousCancellations	Integer	Number of previous bookings that were cancelled by the customer
RequiredCardParkingSpaces	Integer	Number of car parking spaces required by the customer
ReservationStatus	Categorical	Reservation last status
ReservationStatusDate	Date	Date at which the last status was set.
ReservedRoomType	Categorical	Code of room type reserved.
StaysInWeekendNights	Integer	Number of weekend nights (Sat or Sun) the stayed or booked guest to stay at the hotel
StaysInWeekNights	Integer	Number of weeknights (Mon to Fri) the guest stayed or booked to stay at the hotel
TotalOfSpecialRequests	Integer	Number of special requests made by the customer

Data preprocessing and transformation

The sourced data was cleaned and transformed before analysis was carried out. First, the dimension and structure of the data was checked. Then the sapply function was used to convert all the variables in the dataset into numeric. Normalization was done to the numeric data to adjust the values in the dataset to a common scale, without altering the variations in the ranges of values. This was then put into a data frame to make it presentable for the analysis. Using the sapply and is.na function, the data was checked for NA values of which some were discovered. The NAs found were programmatically removed because it did not sum up to 1% of the data.

Data Selection

During data selection, variables not relevant for the analysis were removed. The figure 15 below shows the correlation matrix amongst the variables chosen.

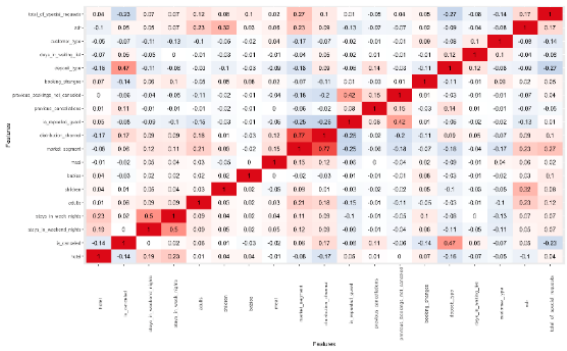

Figure 15: Correlation amongst variable

Implementation of Algorithms

The appropriate libraries were loaded, the data was read in and the dimension and structured of the data was checked to make sure the data was properly entered. The k nearest neighbor(kNN) and support vector machine classifier were used in the hotel booking prediction. The data was split into 70% for training and 30% for testing. The data was trained using the train dataset. After training, it was then tested using the testing dataset for predictions. The summary of the classifiers is shown below.

For the kNN, the k was determined by finding the square root of the total number of observations. There were 83570 observations and the square root gotten was approximately 289.0. Therefore, 289 and 290 where chosen as the k value.

The kNN model was trained fitting k = 289 and 290 and tested. The k= 289 model had an accuracy of 0.8058 which is translated as 81% accuracy. The confusion matrix was used to determine the performance of when k =289 is shown below

	Reference	
Prediction	0	1
0	22020	6384
1	573	6839

This shows that 22020 cases where predicted correctly as not cancelling a booking, 6384 cases who cancelled where predicted as not cancelling, 573 cases who did not cancel a booking where predicted as canceled and lastly, 6839 cases who cancelled were predicted correctly.

The k= 290 model had an accuracy of 0.804 which is translated as 80% accuracy. The confusion matrix was used to determine the performance of when k =290 is shown below

	Reference	
Prediction	0	1
0	22028	6455
1	565	6768

This shows that 22028 cases where predicted correctly as not cancelling a booking, 6455 cases who cancelled where predicted as not cancelling, 565 cases who did not cancel a booking where predicted as canceled and lastly, 6768 cases who cancelled were predicted properly.

The support vector machine was used on the same training and testing data. But unfortunately, it showed a 100% accuracy which is not feasible for a model.

Results

The result shows that the kNN was a better classifier in predicting the possibility of a client canceling a hotel reservation made with an accuracy of 80%.

IV. Evaluation

Evaluating a model is as important as the analysis carried out. The evaluation helps us know how well the model performed and if it will be useful in predicting other problems in the future. The following includes the performance measures that were used in evaluation the different models. The accuracy, ROC curve and AUC, sensitivity, specificity, root squared mean error (RMSE).

The table below gives a summary of all the performance measures and algorithms used.

Credit Card default

Logistics regression				Random forest			
Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC
0.81	0.97	0.22	0.72	0.82	0.84	0.67	0.76

Beijing Air-Quality pollution

PM2.5		NO2		CO	
RMSE	R square	RMSE	R square	RMSE	R square
70.797	0.218	45.980	0.218	1652.974	0.218

Hotel booking

KNN 289	KNN 290	SVM
Accuracy	Accuracy	Accuracy
0.805	0.803	1.000

We can make comparisons between the algorithms implemented from the tables above. The first table was used to classifier credit card users into those likely to default or not. The random forest was a better classifier with an accuracy of 82% and higher area under the curve (AUC) of 0.765 even though it had a higher specificity.

The second table was used to answer which air pollutant had the most influence on the weather condition. Since all the models had the same accuracy of 21%, the RMSE was used to determine which model did not fit the data, thereby equating it to having no influence on the weather conditions provided. Since the CO had the highest RMSE of 1652.9745264, it is regarded that the CO has no relationship with the weather conditions.

Lastly, the k nearest neighbor (kNN) and the support vector machine (SVM) were used to determine the client that will likely cancel on a hotel booking or not. The KNN proved to be a better classifier with an accuracy of 80% than the SVM.

V. Conclusion

In conclusion, the objective of the first analysis was to classify the possibility of a credit card client into defaulting and not. The end result showed that the random forest

classifiers performed better with an accuracy of 82% with a higher area under the curve than the logistics regression.
The next objective was to check the relationship of air pollutant with the weather conditions. Using the multiple linear regression, the data was modeled using three pollutants as the target and the weather conditions as the predictors. The end result shows that NO2 has the most relationship/influence with the weather conditions using multiple linear regression.
The last objective was to check the probability of canceling a hotel booking or not. The final result from the analysis shows that kNN was a better classifier with an accuracy of 80%.

VI. References

[1] Y. B. Wah and I. R. Ibrahim, "Using Data Mining Predictive Models to Classify," in *2010 6th International Conference on Advanced Information Management and Service (IMS)*, South Korea, 2010.

[2] Y. Zhang, Y. Wang, M. Gao, Q. Ma, J. Zhao, R. Zhang, Q. Wang and L. Huang, "A Predictive Data Feature Exploration-Based," *IEEE Access,* vol. 7, pp. 30732-30734, 25 March 2019.

[3] N. Antonio, A. Almeida and L. Nunes, "Hotel booking demand datasets," in *Data in brief*, Lisbon, Portugal, 2019.

[4] M. S. Kumar, . V. Soundarya, S. Kavitha, E. S. Keerthika and E. Aswini, "Credit Card Fraud Detection Using Random Forest Algorithm," in *2019 3rd International Conference on Computing and Communication Technologies (ICCCT)*, Chennai, 2019.

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

[6] A. Subasi and S. Cankurt, "Prediction of default payment of credit card clients using Data Mining Techniques," in *2019 International Engineering Conference (IEC)*, Erbil, Iraq, 2019.

[7] D. Feng, Z. Deng, T. Wang, Y. Liu and L. Xu, "Identification of disturbance sources based on random forest model," in *2018 International Conference on Power System Technology (POWERCON)*, Guangzhou, China, 2018.

[8] S.-T. Luo, B.-W. Cheng and C.-H. Hsieh, "Prediction model building with clustering-launched classification and support vector machines in credit scoring," *Expert Systems with Applications,* vol. 36, no. 4, pp. 7562-7566, May 2009.

[9] L. Peng, L. Siben, B. Tingting and L. Yang, "Telecom Customer Churn Prediction Method Based on Cluster," in *International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things 2014*, Hsinchu, 2014.

[10] A. Z. Ul-Saufie, A. S. Yahya , N. A. Ramli and H. A. Hamid, "Comparison Between Multiple Linear Regression And Feed forward Back propagation Neural Network Models For Predicting PM10 Concentration Level Based On Gaseous And Meteorological Parameters," *International Journal of Applied Science and Technology,* vol. 1, no. 4, pp. 42-49, July 2011.

[11] S. Sousa, F. Martins, M. Alvim-Ferraz and M. Pereira, "Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations," *Environmental Modelling & Software,* vol. 22, no. 1, pp. 97-103, January 2007.

[12] M. Rosenlund, F. Forastiere, M. Stafoggia, . D. Porta, M. Perucci, A. Ranzi, . F. Nussio and C. A. Perucci , "Comparison of regression models with land-use and emissions data to predict the spatial distribution of traffic-related air pollution in Rome," *Journal of Exposure Science & Environmental Epidemiology,* vol. 18, p. 192–199, 2008.

[13] H. Men, S. Zhang, J. Jin and Z. Xu, "Simultaneous Determination of Pb and Cd Ions with Ion Selective Electrodes Based on Multiple Linear Regression," in *2009 Third International Symposium on Intelligent Information Technology Application*, Shanghai, China, 2009.

[14] D. Wei, M. Xing, J. Zhang, C. Zhang and . H. Cao, "Applied Research of Multiple Linear Regression in the Information Quantification of Chinese Medicine Bone-setting Manipulation," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Madrid, Spain,, 2008.

[15] S. W. YouLi Feng, "A Forcast for Bicycle Rental Demand Based on Random Forests and Multiple Linear Regression," 2017.

[16] T. M. Amado and J. C. Dela Cruz, "Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization," in *TENCON 2018 - 2018 IEEE Region 10 Conference*, Jeju, Korea (South), 2018.

[17] H.-C. Huang, . A. Y. Chang and . C.-C. Ho, "Using Artificial Neural Networks to Establish a Customer-cancellation Prediction Model," pp. 178-180, January 2013.

[18] N. Antonio, A. M. De Almeida and L. Nunes, "An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations," *Data Science Journal,* vol. 18, no. 32, p. 1–20, 2019.

[19] C.-C. Chen, Z. Schwartz and P. Vargas, "The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers," *International Journal of Hospitality Management,* vol. 30, no. 1, pp. 129-135, March 2011.

[20] D. R. Moral and J. Wang, "Forecasting cancellation rates for services booking revenue management using data mining," *European Journal of Operational Research,* vol. 202, no. 2, pp. 554-562, 16 April 2010.

[21] A. Ansari, A. Shaikh, S. Mapkar and M. Khan, "Cancellation Prediction for Flight Data Using Machine Learning," *2nd International Conference on Advances in Science & Technology (ICAST) 2019 on 8th, 9th April 2019 by K J Somaiya Institute of Engineering & Information Technology,* 2019.