Paper 3471-2015

Forecasting Vehicle Sharing Demand using SAS® Forecast Studio

Kushal Kathed, Ayush Priyadarshi and Dr. Goutam Chakraborty, Oklahoma State University

ABSTRACT

As pollution and population continue to increase, new concepts of ecofriendly commutating evolve. One of the emerging concepts is the bicycle sharing system; it is a bike rental service on a short-term basis at a moderate price. It provides people the flexibility to rent a bike from one location and return it to another location. This business is quickly gaining popularity all over the globe.

In May 2011, there were only 375 bike rental schemes comprised of nearly 236,000 bikes in USA. However, this number jumped to 535 bike sharing programs with approximately 517,000 bikes in just a couple of years. It is expected that this trend will continue to grow at a similar pace in the future.

Most of the businesses involved in this system of bike rental are faced with the challenge of balancing between supply and inconsistent demand. The number of bikes needed on a particular day can vary on several factors such as season, time, temperature, wind speed, humidity, holiday and the day of the week. In this paper we have illustrated an application of SAS® Forecast Studio to attempt to solve the forecasting problem. Incorporating the effect of all the above factors and analyzing the demand trends of the last two years, we have been able to precisely forecast the number of bikes needed on any day in the future. From a managerial perspective, we are also able to illustrate the scenario analysis to observe the effect of particular variables on demand.

INTRODUCTION

Forecasting the number of bike rentals on the next day or over the next week can be a difficult task for the managers because of its erratic demand. In this paper, we analyze two years data of a bike rental company in Washington D.C to:

- Analyze each independent variable in relation to the target to figure out the most critical attributes that increase/decrease bike rental demand.
- Build a time series predictive model to predict the total number of renters.
- Provide the manager with the forecasted value of the bike demand for each day of the next to make appropriate business decisions.
- Perform scenario analysis and observe the change in the forecasted values by changing the values of weather and temperature on a particular day.

DATA DICTIONARY

Variable Name	Level	Description
Season	Nominal	This field represents the type of season. 1= Spring, 2 = Summer, 3 = Fall, 4 = Winter
Year	Nominal	This field corresponds to the fiscal year. 0 = 2011, 1 = 2012
Month	Nominal	This represents the month. 1 = January, … , 12 = December
Holiday	Nominal	This variable represents whether it is holiday or not. 1 = Holiday, 0 = Not Holiday
Weekday	Nominal	This indicates the weekday. 0 = Sunday, … , 6 =Saturday

Working_Day	Nominal	This field indicates whether it is working day or not. 1 = Working day, 0 = Not working day
Weather_Situation	Nominal	This field represents the type of weather situation. 1= Clear, 2 =Cloudy, 3 = Light Snow/Rain 4 = Heavy Snow/Rain
Temperature	Interval	This field represents the temperature in Fahrenheit.
Feel_Temperature	Interval	This represents the feel temperature in Fahrenheit.
Humidity	Interval	It represents the humidity in percentage (relative humidity).
Wind_Speed	Interval	It corresponds to the wind speed in mph.
Casual_Renters	Interval	This indicates the count of casual renters.
Registered_Renters	Interval	This field represents the count of registeredrenters.
Renters_Count	Interval	This field provides the count of total renters.

Table 1. Data dictionary for bike sharing data

DATA PREPARATION

The data for our analysis was downloaded from UCI Machine Learning repository. The dataset comprised of twelve input variables such as season, date, temperature, wind speed, humidity and holiday. The data consists of 731 bike rental records starting from Jan 01, 2011 till Dec 31, 2012. The last three months of the data has been used as hold out sample for validating our model whereas rest of the data is used to train the model. Using the training and validation data, we are able to come up with the forecasted value of total number of bike renters for each day of the next year starting Jan 01, 2013 with 95 percent confidence interval. For the purpose of analysis, new dummy variables were created for all the categorical variables.

OUTLIER DETECTION

Parameter Estimates

Component	Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
AO24MAR2012D	SCALE	-1349.0	257.71157	-5.23	<.0001
AO07OCT2012D	SCALE	-1377.9	256.59400	-5.37	<.0001
AO04JUL2011D	SCALE	1880.4	255.29076	7.37	<.0001
AO17MAR2012D	SCALE	1146.5	257.78028	4.45	<.0001
AO27AUG2011D	SCALE	-1418.9	255.48518	-5.55	<.0001
AO28APR2012D	SCALE	-1070.2	256.11731	-4.18	<.0001
AO04JUL2012D	SCALE	1462.3	255.50264	5.72	<.0001
AO22APR2012D	SCALE	-1265.4	264.28224	-4.79	<.0001
AO01JUN2012D	SCALE	-1170.2	255.63047	-4.58	<.0001

Table 2. Outliers in the data

SAS® Forecast studio has the capability to automatically detect outliers in the data based on the significance level. Table 2 shows the 9 outliers in our data detected using SAS® Forecast studio. We further tried to study some of these outliers and discovered some interesting results.

Consider two of the outlier dates:

- Aug 27, 2011 (Hurricane Irene)
- July 04, 2011 (Independence Day)

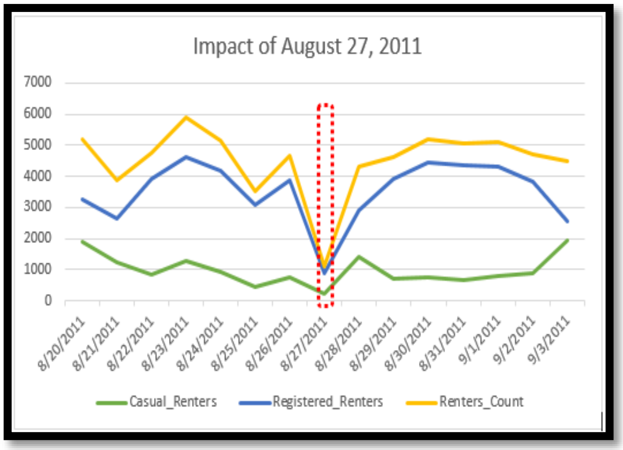

Figure 1. Variation due to outlier

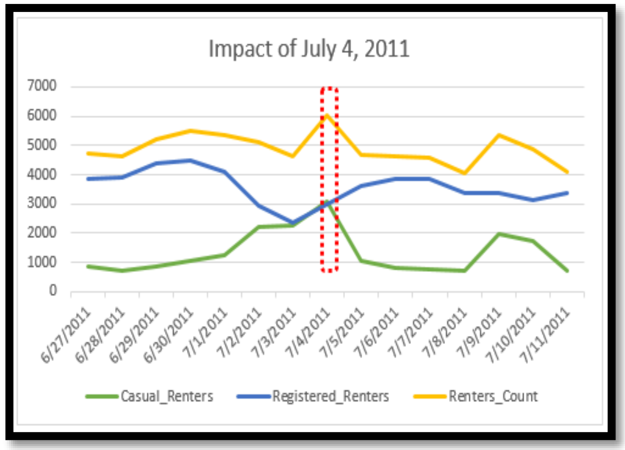

Figure 2. Variation due to outlier

The above two figures delineates the number of casual renters, registered renters and total bike renters on the two outlier dates. August 27, 2011 experienced a sudden decrease in the number of bike renters for all the renter categories as on that day Hurricane Irene impacted Washington D.C area and all means of transport were aborted. On the other hand, on July 04, 2011which is also the Independence Day, there is a sudden rise in number of casual renters thereby also increasing the total number of renters. This increase is expected as most of the casual renters on that date would be tourists visiting the city.

TIME SERIES ANALYSIS

Time series model learns the trends, seasonality and cyclicality in the data and forecast the future values. These models are built in SAS/ETS (Econometrics and Time Series) software by using the SAS® code below.

Building time series models such as ARIMA in SAS/ETS, requires AR (autoregressive) and MA (moving average) values. To figure out the best combination of AR and MA, analysts will have to manually evaluate all the models built with different AR and MA combinations. Here as an example we have used AR (p) =1 and MA (q) =1.

```
/* SAS® code to build ARIMA model */
Ods graphics on/imagemap=on;
Proc ARIMA data=COURSE.DailyBikeData plots(unpack)=series(all);
Identify var=Renters_Count(1, 4) nlags=12;
estimate p=1 q=1;
forecast lead=360 interval=daily id=date out=results;
run;
ods graphics off;
```

If data has seasonality it may be best to remove seasonality from the data. For this purpose we used X12 procedure which captures seasonality effect in a variable. These seasonality effects can be applied back later to the forecasted values once the model has been built. X12 is one of the widely used procedures to remove seasonality from the data. Below code demonstrates the use of X12 procedures in SAS/ETS.

```
/* seasonal adjustment using proc X12 from SAS */
Ods graphics on/imagemap=on;
Proc X12 data=course.DailyBikeData date=date;
Var Renters_Count;
x11;
outputout = out a1 d10 d11 d12 d13;
run;
ods graphics off;
```

VARMAX procedure provides the flexibility to find out the best combination of AR and MA among multiple AR and MA combinations. Below code demonstrates the use of VARMAX to find best ARIMA model.

```
/* Running VARMAX on seasonally adjusted data */
Ods graphics on/imagemap=on;
Proc VARMAX data=out plot=all;
id date interval=day;
model Renters_Count_D11/  method=ml lagmax=12;
output out=out lead=360;
run;
ods graphics off;
```

There are other time series models available such as Exponential Smoothing Models (ESM), Unobserved Component Model (UCM) and so on. Sometimes, ESM model performs better than ARIMA model. So to figure out the best time series model we need to run multiple models and compare them on the basis of fit statistics such as Mean Absolute Percent Error (MAPE).

Performing all of the above tasks for determining the best time series model can be a tedious task. SAS® Forecast Studio provides us the capability to quickly perform all of the above tasks as well as others required to analyze and diagnose the time series model.

SAS® Forecast Studio is a user-friendly interface to perform time series analysis. It provides the power to run multiple time series models and find the best one among them. As shown below, it also has the functionality to automatically detecting and removing outliers, perform scenario analysis which could be used by managers to visualize the impact of a predictor variable on the target variable.

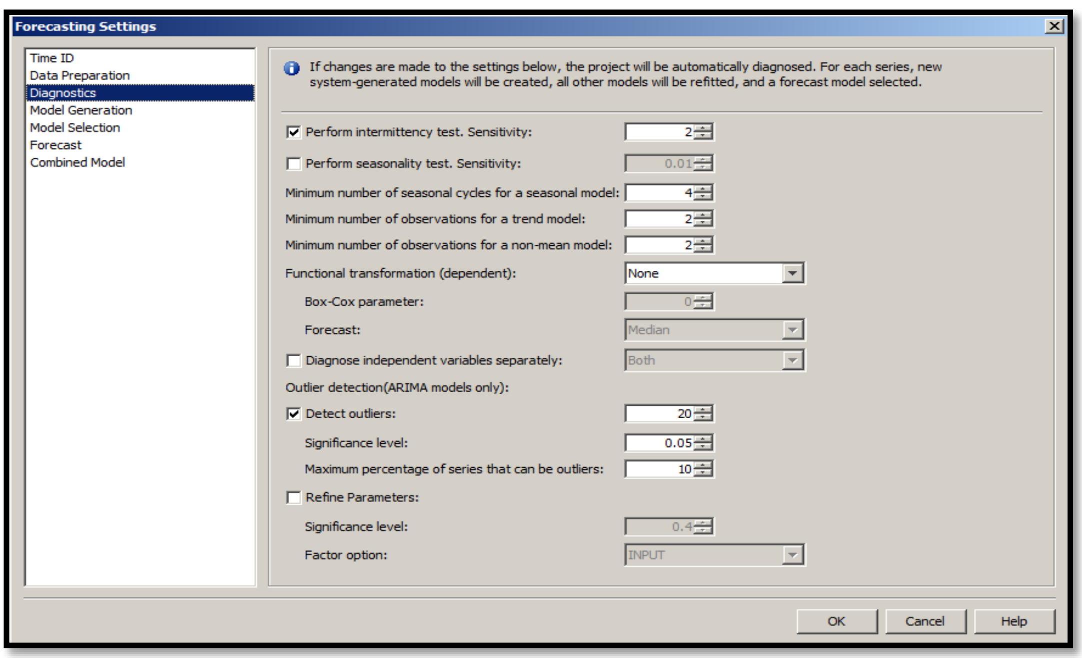

Figure 3. SAS® Forecast Studio settings window

In order to build a time series model, we used first 21 months (Jan 01, 2011 to Sept 30, 2012) of data to train the model and last 3 months (Oct 01, 2012 to Dec 31, 2012) of data to validate the model. The best model is used to forecast the total bike requirement for next year (Jan 01, 2013 to Dec 31, 2013). Below figure depicts that model has learnt the trend and seasonality pretty well from the data.

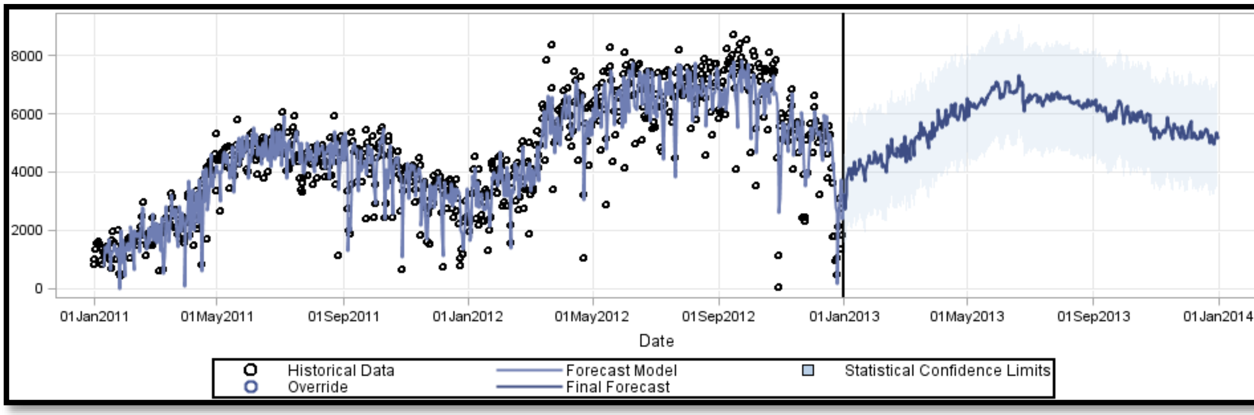

Figure 4. Time series for total renters

ARIMA model turned out to be the best model for this data. Below figure shows the significant variables in the selected model and their respective parameter estimate values.

Parameter Estimates

Component	Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Renters_Count	CONSTANT	-3094.9	759.08352	-4.08	<.0001
Renters_Count	MA1_1	0.29101	0.06343	4.59	<.0001
Renters_Count	AR1_1	0.75368	0.04458	16.91	<.0001
Temperature	SCALE	66.80648	7.56103	8.84	<.0001
Temperature	NUM1_1	20.77509	7.95016	2.61	0.0092
Temperature	NUM1_2	-9.55053	7.14454	-1.34	0.1817
WindSpeed	SCALE	-28.87101	5.85372	-4.93	<.0001
WindSpeed	DEN1_1	0.27623	0.20630	1.34	0.1810
Weather_Clear	SCALE	2127.4	189.34416	11.24	<.0001
Weather_Clear	NUM1_1	-323.95031	184.80062	-1.75	0.0800
Weather_Clear	NUM1_2	15.07105	177.85409	0.08	0.9325
Weather_Cloudy	SCALE	1524.9	189.89428	8.03	<.0001
Weather_Cloudy	NUM1_1	-366.16774	187.55288	-1.95	0.0513
Weather_Cloudy	NUM1_2	-21.22127	177.89256	-0.12	0.9051
Weather_RainOrSnow	SCALE	-234.02868	99.29430	-2.36	0.0187
Weather_RainOrSnow	DEN1_1	0.98719	0.0074637	132.26	<.0001
Weekday_Wednesday	SCALE	139.85749	53.16814	2.63	0.0087
Weekday_Wednesday	DEN1_1	0.30184	0.30413	0.99	0.3213
Weekday_Wednesday	DEN1_2	0.69398	0.30425	2.28	0.0229
Season_Fall	SCALE	267.83363	368.67404	0.73	0.0092
Season_Fall	NUM1_1	508.69895	368.40325	1.38	0.1678
Season_Summer	SCALE	-433.10651	374.90207	-1.16	<.0001
Season_Summer	NUM1_1	-600.30985	428.26422	-1.40	0.1614
Season_Summer	NUM1_2	-318.21345	373.24659	-0.85	0.3942

Table 3. Parameter estimates for total renters

MODEL COMPARISON

SAS® Forecast Studio provides an useful feature to compare the generated models visually. In situations if one model is picking up the trend and other is picking up the seasonality then we could merge both the models and create a combined model consisting properties of both the models.

Model	Type	Read-Only	Holdout MAPE
Generated ARIMA Model (LEAF_1)	**Generated**	**Yes**	**31.16**
Generated Smoothing Model (LEAF_2)	Generated	Yes	85.23
Generated ARIMA Model (LEAF_0)	Generated	Yes	180

Table 4. Model comparison for total renters

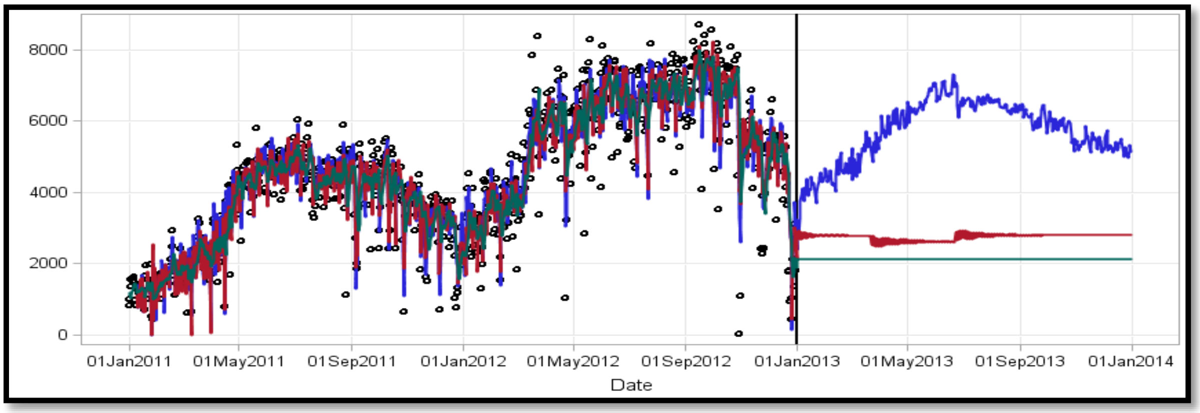

Figure 5. Visual model comparison

For model evaluation we are considering the Mean Absolute Percentage Error (MAPE) values.

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Where, M= MAPE, A_t= Actual value, F_t= Forecasted value, n = no. of fitted points

The model selected for forecasting the total number of bike renters is the ARIMA model with accuracy of 69%.

	Casual Renters	Registered Renters	Total Renters
Best Model	ARIMA	ARIMA	ARIMA
MAPE	45	37	31
R-Square	0.79	0.88	0.85
RMSE	235.5	531.49	746.5

Table 5. Model accuracy comparison

SCENARIO ANALYSIS

Another extremely useful feature of SAS® Forecast Studio is the ease of performing Scenario Analysis which allows a manager to create what-if scenarios based on the selected model. For example, we now know the expected number of bike renters on a particular day in future based on our model, but the manger may be interested in knowing the impact of sudden change in weather on the bike demand. For these cases SAS® Forecast Studio has an inbuilt Scenario Analysis feature where we can manually change the values of our significant input variables and thereby observe its effect on the target variable. In other words, it provides the flexibility to vary the independent variables to see their impact on the target variable. Let us look at some of the scenario analysis using the tool.

Scenario I: Weather is changed from Cloudy to Clear on June 02, 2013

Impact on target: This change leads to bike demand increase by approximately 1,000 units.

	02Jun2013
* TEMPERATURE	70
* WINDSPEED	13
* WEATHER_CLEAR	1
* WEATHER_CLOUDY	0
* WEATHER_RAINORSNOW	0
* WEEKDAY_WEDNESDAY	0
* SEASON_FALL	0
* SEASON_SUMMER	1
Model Forecast	6,527
Scenario Forecast	7,599

Table 6. Weather change Scenario

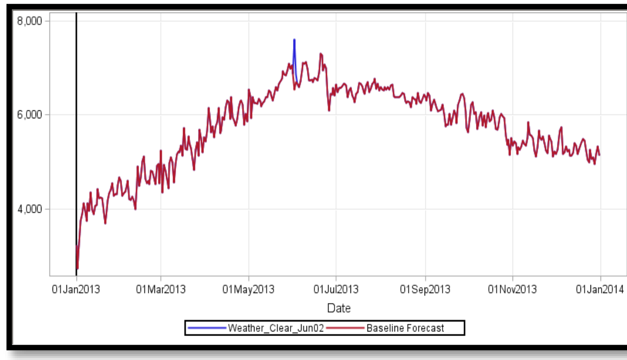

Figure 6. Effect on predicted values

Scenario II: Wind speed increased from 10 mph to 23 mph on July 19, 2013,

Impact on target: Bike rental demand goes down by 300 units.

	19Jul2013
* TEMPERATURE	82
* WINDSPEED	23
* WEATHER_CLEAR	0
* WEATHER_CLOUDY	0
* WEATHER_RAINORSNOW	0
* WEEKDAY_WEDNESDAY	0
* SEASON_FALL	1
* SEASON_SUMMER	0
Model Forecast	6,613
Scenario Forecast	6,321

Table 7. Weather change Scenario

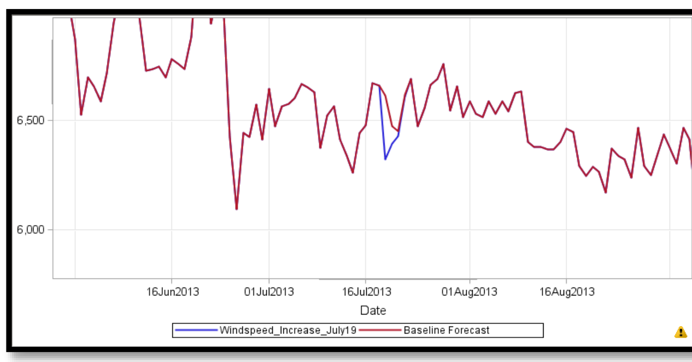

Figure 7. Effect on predicted values

DISCUSSION

Thus, using SAS® forecast Studio we were able to forecast the bike rental demand for each day of the next year considering all the factors. Observing the forecasted values, managers can make appropriate decision on number of bikes that would be required on a particular day.

Below table displays the forecasted values of bike demand for the first week of January in 2013 with 95 percent confidence interval as predicted by the model.

Forecast Values

Obs	Date	Forecasts	Standard Error	LOWER	UPPER
732	01JAN2013	3218.7539	759.2574	1730.6367	4706.8712
733	02JAN2013	2743.9026	836.5836	1104.2288	4383.5764
734	03JAN2013	3268.4451	877.4776	1548.6206	4988.2695
735	04JAN2013	3744.8570	899.8792	1981.1261	5508.5878
736	05JAN2013	3889.3360	912.3590	2101.1452	5677.5268
737	06JAN2013	4115.7140	919.3724	2313.7772	5917.6509
738	07JAN2013	3958.9915	923.3325	2149.2931	5768.6900

Table 8. Forecasted values for next year

In the final model, we have identified the following significant factors affecting the bike rental demands

- Temperature
- Clear weather
- Cloudy weather
- Rain/Snow weather
- Summer season
- "Wednesday??"

Rise in temperature, presence of clear weather and summer season generally drive the bike rental business whereas cloudy weather and Rain/Snow weather decreases number of renters. Although we couldn't understand the significance of the weekday "Wednesday" in our model at this point due to limitations with the size of data but this is something really interesting and perhaps some domain expertise would reveal why Wednesday is showing higher sales.

The scenario analysis feature inside SAS® Forecast Studio a great tool for the managers to make instant decision to slash or increase the tariff rates based on the weather forecast to boost their revenue.

REFERENCES

1) Introducing SAS® Forecast Studio by Brenda Wolfe, Michael Leonard, Paddy Fahey(SAS Institute Inc., Cary, NC)
2) http://www.sas.com/en_us/software/analytics/forecastserver.html
3) http://www.sas.com/resources/whitepaper/wp_3495.pdf
4) Forecasting Enrollment in Higher Education using SAS® Forecast Studio® by Erik Bowe, Steven Merritt (Kennesaw State University)

ACKNOWLEDGMENTS

This paper utilized data from the UC Irvine Machine Learning Repository. We are thankful to them for providing us the data for analysis purpose.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kushal Kathed
Oklahoma State University
Stillwater, OK, 74075
Work Phone: (405)762-2880
Email: kushal.kathed@okstate.edu
Kushal Kathed is a graduate student enrolled in Management Information Systems at Spears School of Business, Oklahoma State University (OSU), Stillwater. He is very passionate about analytics and want to pursue his career in this field. He is working as a Research Assistant on a consulting project for a leading gas company in Oklahoma. He completed his internship at BNSF Railway, Fort Worth in summer '14. He is an Advanced SAS, Statistical Business Analyst certified, also holds the SAS and OSU Data Mining Certificate. He has presented two posters at the SAS Analytics Conference 2014.

Ayush Priyadarshi,
Oklahoma State University
Stillwater, OK, 74075
Work Phone: (405)780-5298
Email: ayushp@okstate.edu
Ayush Priyadarshi is a Graduate student enrolled in Management Information Systems at Spears School of Business, Oklahoma State University (OSU), Stillwater. He has more than three years work experience in data analytics and Information Technology. He is a Base SAS® 9 certified professional, a certified SAS Statistical Business Analyst, JMP Software data exploration certified and holds the SAS and OSU Data Mining certificate. He has previously given three poster presentations at the SAS Analytics Conference 2014 and is also the author of another paper at SAS Global Forum 2015 named 'Using text from Repair Tickets of a Truck Manufacturing Company to predict factors that contribute to truck downtime'.

Dr. Goutam Chakraborty
Oklahoma State University
goutam.chakraborty@okstate.edu
Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS and OSU data mining certificate and SAS and OSU marketing analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has over 25 Years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.