/\*

Dhaval Delvadia

CSC 423 Date 3/12/2017

Final Project - Predicting Hourly Bike Rentals

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors. The core data set is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is

publicly available in http://capitalbikeshare.com/system-data. We aggregated the data on two hourly and daily basis and then extracted and added the corresponding weather and seasonal information. Weather information are extracted from http://www.freemeteo.com.

\*/

\* Uploading the bike sharing data to SAS;

**DATA** bike\_sharing;

INFILE "C:\Users\dhaval&mitali\Documents\CSC 423\Proj\Bike-Sharing-Dataset\Code\data.csv" delimiter=',' firstobs=**2**;

INPUT instant dteday :mmddyy10. season yr mnth hr holiday weekday workingday weathersit temp atemp hum windspeed casual registered cnt;

format dteday mmddyy10.;

**Run**;

\*Setting day, raw\_temp, raw\_atemp, raw\_hum, raw\_windspeed, dummy seasons variables, dummy weather variables;

**DATA** bike\_sharing;

SET bike\_sharing;

day=DAY(dteday);

raw\_temp=(temp\***41**);

raw\_atemp=(atemp\***50**);

raw\_hum=(hum\***100**);

raw\_windspeed=(windspeed\***67**);

d\_summer=(season=**2**);

d\_fall=(season=**3**);

d\_winter=(season=**4**);

d\_mist=(weathersit=**2**);

d\_LightSnow=(weathersit=**3**);

d\_heavyRain=(weathersit=**4**);

/\* peak hours for bike rentals between 7-9 am and 5-7 pm. On weekends, the peak hours were anywhere between 10 am - 6 pm\*/;

if workingday = **1** and (**7** <= hr <= **9** or **17** <= hr <=**19**) then week\_peak =**1**;

else week\_peak= **0**;

if workingday = **0** and (**10** <= hr <= **18**) then weekend\_peak = **1**;

else weekend\_peak=**0**;

**Run**;

\*Creates a new training and testing dataset, xt\_bs, adds a column "select" splitting train and test sets;

**PROC** **SURVEYSELECT** DATA=bike\_sharing

OUT = xt\_bs SEED=**91558**

SAMPRATE = **0.7** OUTALL;

title "Test and Train Sets for Bike Sharing Demand";

**RUN**;

\*Drawing histogram of cnt;

**PROC** **UNIVARIATE** data=xt\_bs;

VAR cnt;

histogram /normal (mu=est sigma=est);

**run**;

\* create new variable new\_sqrt\_cnt for training set, and = NA for testing set;

**data** xt\_bs;

set xt\_bs;

sqrt\_cnt = sqrt(cnt);

log\_cnt=log(cnt);

Inv\_cnt = **1**/cnt;

sqr\_cnt=cnt\*\***2**;

cube\_cnt=cnt\*\***3**;

if selected then new\_y=sqrt\_cnt;

**run**;

\*Drawing histogram of new\_sqrt\_cnt, and etc;

**PROC** **UNIVARIATE** data=xt\_bs;

VAR log\_cnt Inv\_cnt sqr\_cnt cube\_cnt;

histogram /normal (mu=est sigma=est);

**run**;

\*Draw scatter plots of indipendent vs dependent variables with square root transformation of dependent variable cnt;

**PROC** **GPLOT**;

PLOT new\_y\*hr new\_y\*day new\_y\*mnth new\_y\*yr new\_y\*d\_summer new\_y\*d\_fall

new\_y\*d\_winter

new\_y\*holiday new\_y\*weekday new\_y\*workingday new\_y\*d\_mist new\_y\*d\_LightSnow

new\_y\*d\_heavyRain new\_y\*raw\_temp new\_y\*raw\_atemp new\_y\*raw\_hum new\_y\*raw\_windspeed

new\_y\*week\_peak

new\_y\*weekend\_peak;

**RUN**;

\*Matrix plot;

**PROC** **SGSCATTER**;

\*MATRIX new\_y raw\_temp raw\_atemp raw\_hum raw\_windspeed;

matrix new\_y day mnth yr d\_summer d\_fall d\_winter

holiday weekday workingday d\_mist d\_LightSnow

d\_heavyRain raw\_temp raw\_atemp raw\_hum raw\_windspeed

week\_peak weekend\_peak;

**RUN**;

\*Correlation with transformed dependent variable cnt as square root of cnt to find multicolliniarity;

**PROC** **CORR**;

VAR new\_y hr day mnth yr d\_summer d\_fall d\_winter

holiday weekday workingday d\_mist d\_LightSnow

d\_heavyRain raw\_temp raw\_atemp raw\_hum raw\_windspeed

week\_peak weekend\_peak;

**RUN**;

\*Model 1 includes all dep Var and from that model we find all the variables with variance inflation factor/multicolliniarity and p-val to remove null hypothesis variables;

**proc** **reg** data=xt\_bs PLOTS (MAXPOINTS = NONE);

title "Model 1";

model new\_y = hr day mnth yr d\_summer d\_fall d\_winter

holiday weekday workingday d\_mist d\_LightSnow

d\_heavyRain raw\_temp raw\_atemp raw\_hum raw\_windspeed / VIF TOL stb;

**run**;

\*Model 2 without raw\_windspeed, d\_heavyRain, and day since p-val higher than 0.05

and without raw\_temp var since VIF > 10;

**proc** **reg** data=xt\_bs PLOTS (MAXPOINTS = NONE);

title "Model 2";

model new\_y = hr yr d\_summer d\_fall d\_winter

holiday weekday workingday d\_mist d\_LightSnow

raw\_atemp raw\_hum

week\_peak weekend\_peak / VIF TOL stb;

\*output out = xt\_bs rstudent=r cookd=cd; \*outlier and influent point detected;

**run**;

\*Model 3 to find out what are the best variables from model 2

by employing stepwise, adj-r2, and cp selection methods;

**proc** **reg** data=xt\_bs PLOTS (MAXPOINTS = NONE);

title "Model 3";

model new\_y = hr yr d\_summer d\_fall d\_winter

holiday weekday workingday d\_mist d\_LightSnow

raw\_atemp raw\_hum

week\_peak weekend\_peak / selection=stepwise sle=**0.05** sls=**0.05**;

model new\_y = hr yr d\_summer d\_fall d\_winter

holiday weekday workingday d\_mist d\_LightSnow

raw\_atemp raw\_hum

week\_peak weekend\_peak / selection=adjrsq cp;

**run**;

\* get predicted values for the missing new\_y in test set for 2models;

**proc** **reg** data=xt\_bs PLOTS (MAXPOINTS = NONE);

title "Validation - Test Set";

\* MODEL 3;

model new\_y = hr yr d\_summer d\_fall d\_winter

holiday weekday workingday d\_mist d\_LightSnow

raw\_atemp raw\_hum

week\_peak weekend\_peak;

\*out=outm1 defines dataset containing Model 1 predicted values for test set;

output out=outm1(where=(new\_y=**.**)) p=yhat;

**run**;

/\* summarize the results of the cross-validations for model-1\*/

title "Difference between Observed and Predicted in Test Set";

**data** outm1\_sum;

set outm1;

d=sqrt\_cnt-yhat; \*d is the difference between observed and predicted values in test set;

absd=abs(d);

**run**;

/\* computes predictive statistics: root mean square error (rmse)and mean absolute error (mae)\*/

**proc** **summary** data=outm1\_sum;

var d absd;

output out=outm1\_stats std(d)=rmse mean(absd)=mae ;

**run**;

**proc** **print** data=outm1\_stats;

title 'Validation statistics for Model';

**run**;

\*computes correlation of observed and predicted values in test set;

**proc** **corr** data=outm1;

var sqrt\_cnt yhat;

**run**;