# Predictive Analytics

#### **Part One—Identify a Focal Point and a Dependent Variable**

1. Briefly summarizing the decision or prediction that we wish to target using predictive analysis

Target of analysis:

|  |
| --- |
| With the collected UCI bike share data, we are about to predict the bike rentals demand and to start new bike stations in new places with the available variables. The total process from membership, renting the bikes and returning them to a nearby stand is completely automatic. Traffic, Environmental conditions and health risks in everyday life will increase the trend in renting bicycles. Here, we have the dataset with years 2011 and 2012. With this we are about to predict the demand in bikes respective to weekday, holiday, weather and temperature conditions, season, etc. Users can rent a bike with paid membership or as a casual user. We also have seasonal, weather and temperature information. Our prediction is on number of bikes rented. So, our dependent variable for this analysis is BikesCount variable. |

Description of the context:

|  |
| --- |
| During traffic jams at peak working hours, people are made to wait for hours. But easier in renting a bike to travel places during this crisis hours is a new trend in the city. It also helps people to avoid health risks. People show great interest in this bike sharing systems. As the time and duration of travel are recorded accordingly, we are able to know the requirement of bikes. In addition to hour and daily count, we also have the corresponding weather information. It helps us to study and analyze in a better way. We will be using multiple regression outputs and correlation matrix table to find the best predictors for our decision. |

1. Dependent variable that guides our prediction or decision.

Dependent variable:

|  |
| --- |
| Bike Count |

1. Identification of variables

|  |  |  |
| --- | --- | --- |
| **Independent Variable** | | |
| **Summary of independent variable** | **Categorical or quantitative?** | **Argument for / description of the associates with the dependent variable** |
| Season | Categorical | 1. Winter 2. Spring 3. Summer 4. Fall   Are the four categories under Season variable. Season might have a good association with the number of bikes rented. |
| Weather | Categorical | 1. Clear 2. Mist and cloudy 3. Light snow 4. Heavy Rain   It is obvious that during Heavy rain people won’t prefer renting a bicycle. During clear sky many people prefer renting bicycles. Thus, weather could be a good predictor of the bike rental demand. |
| Month | Categorical | Month 1-12  We can split the months based on seasons also. There will be some strong collinearity between season and month. So, this is also associated with our dependent variable. |
| Holiday | Categorical | 0 – No Holiday 1 – Holiday  If it is not a holiday, then it is a working day. Working professionals choose to rent bikes during working days and some casual users prefer bikes during weekends for fun and entertainment. It might be a good predictor but we also have working day as another variable. |
| Weekday | Categorical | Weekday 0-Sunday 6-Saturday  With this variable we can find the demand for every day in a week. This is also a good predictor for our dependent variable |
| Working Day | Categorical | 0 -Either a weekend or a holiday  1- Neither a weekend nor a holiday  Working day is going to be a best predictor of our dependent variable. Based on this variable, we can analyze the trend with weekday. They will be good to find with Monday-Friday working hours. |
| Temperature | Quantitative | Temperature is a continuous variable and we can analyze even better with this variable. We can see the preference of bike rentals over certain low or medium or high temperatures. Temperature has a better association with the dependent variable. |
| Atemp | Quantitative | Atemp is feel-like temperature, which has values similar with respect to the temperature variable. Thus, they will have similar association as temperature with the dependent variable. |
| Windspeed | Quantitative | With further analyses we can find the association with the dependent variable. But higher windspeed reduces the renting of bikes. Thus, they might have some influence in the bike rental demand. |
| Humidity | Quantitative | Humidity it the amount of water in air. Windspeed and humidity seems to be related. Thus, they also have some association with our dependent variable. |

**Part Two—Map Decisions to Outcomes**

Mapping how our decision or predicted outcome is related to the independent variables we are considering.

|  |  |  |
| --- | --- | --- |
| **Candidate Independent variables** | | |
| **Independent variable** | **Regression equation** | **Screenshot of scatterplot** |
| Season  1-Winter  2-Spring  3-Summer  4-Fall |  |  |
| Weather  1-clear  2-mist  3-light snow  4-heavy rain |  |  |
| Month  (1-12) |  |  |
| Holiday  0-No  1-Yes |  |  |
| Weekday (0- Sunday, 6- Saturday) |  |  |
| Working Day  (0-No,  1-Yes) |  |  |
| Temperature | y = 6640.7x + 1214.6 |  |
| Atemp | y = 7501.8x + 945.82 |  |
| Windspeed | y = -5862.9x + 5621.2 |  |
| Humidity | y = -1369.1x + 5364 |  |

1. **For each** of the independent variables listed in part 1, explaining why the relationship shown in the scatterplot and regression equation make sense in the context of the situation we are exploring.

|  |  |
| --- | --- |
| Season | It is seen, the range of bike rentals in spring season is less compared to the other 3 seasons. This is a good predictor for our dependent variable. They have a positive association and some linearity. On comparing season vs temperature, mean temperature of spring is low and fall is high. Also, the coefficient value falls between the 95%CI upper and lower bound and are significant enough. P-value <0.05 indicates that they have a significant relationship with the dependent variable. |
| Weather | Weather highly depends on Season. Thus, they also have a positive relationship with the bike rentals. We don’t have significant records on Heavy rain conditions. It is obvious that people don’t prefer bikes during heavy rain conditions. Weather under clear condition is a highly positive predictor of bike rentals. P-value is <0.05 and we reject the null hypothesis frames that there is no relationship with the dependent variable. Thus, they have a significant relationship with the dependent variable. But comparing the absolute value of t-stat -0.674021305 this seems to be statistically insignificant. Also, the coefficient value falls between the 95%CI upper and lower bound and are significant enough. |
| Month | Depending on the season and weather conditions, bike rentals differ for months. With the scatterplot month is a good predictor of bike rentals. It seems season and month will have high correlation between them. Weather and months are highly correlated. |
| Holiday | With the scatter plot it seems, bike rentals are high when it is not a holiday. Which means bike rentals is high in a working day. This is evident that Holiday might not be a good predictor. The negative slope indicates that they don’t a significant relationship. P-value is also >0.05 thus stating insignificant relationship with the dependent variable. Even though the coefficient value falls between the 95%CI upper and lower bound and are significant, we will be eliminating this variable in our model. |
| Weekday | Average rental of bikes rented during the weekdays is same for all the days. Saturday and Sunday also have high rental of bikes where we can assume that for fun and entertainment. The absolute t-stat value indicated that the weekday is significant enough to explain the variation in the dependent variable. Also, p-value <0.05 indicates that there is a relationship between them. Comparing the coefficients, working day shows high positive effect on the bike rentals than the weekday. But we will be elimination this variable from our model later. |
| Working Day | Working professionals tend to use bikes at a higher rate. Correlation coefficient 0.06 shows that it has a strong direct connection with the dependent variable. But absolute t-stat value and p-value >0.05 indicates that there is no relationship between them. But they are a good predictor of the bike rentals demand. |
| Temperature | Slope is positive and with R-squared value we can explain39% of the variation in the dependent variable. They have a highly positive correlation with the bike rentals demand. Higher the temperature, the rate of bike rentals also increases. They are also a good predictor of the dependent variable. They seem to have linearity. Absolute t-stat value and p- value <0.05 indicated that they have a statistically significant relationship with the dependent variable. |
| Atemp | Temperature and Atemp variables show same variations. They tend to show some high multicollinearity. Atemp is a feel like temperature and is formed with respect to the temperature values. Absolute t-stat value and p- value <0.05 indicated that they have a statistically significant relationship with the dependent variable. |
| Windspeed | Windspeed has a negative association with the dependent variable. With high windspeed, the bike rentals count is less. For every unit increase in the windspeed the rental demand decreases by 5862 numbers. Also, they explain only 5% of the variation in the dependent variable. It might not be a good predictor of the bike rentals demand. |
| Humidity | Humidity has a negative correlation/association with the dependent variable. It just explains 1.0% of variation in the dependent variable. With high humidity, the bike rentals count is less. Slope is negative. It might not be a good predictor of the dependent variable. We will be eliminating this variable in our model. |

1. Multiple regression

****

**Part ThreeGenerate a Revised Regression Equation**

Refining the regression model potentially by transforming variables for nonlinear relationships and/or including or excluding variables from the regression to address multicollinearity.

Addressing nonlinear relationships:

1. **For each** of the independent variables listed in part 1, following are the residual plot against the dependent variable Y.





1. Reviewing the residual plots and listing suspected nonlinear variables in the table below.

|  |  |
| --- | --- |
| Temperature | Variance of the residuals are increasing with x. This shows non-linearity |
| Atemp | Similar to temperature plot. This also shows some non-linearity |
| Windspeed | This shows a pattern and thus it has some non-linearity with the dependent variable. |
| Humidity | This shows a pattern and has some non-linearity with the dependent variable. |

1. Using the Semilog and Log-log Transforms tool to create a transform scatterplot for each independent variable listed.

Listing the independent variables with nonlinear relationships.

|  |  |  |
| --- | --- | --- |
| **Possible nonlinearities** | | |
| **Independent variable** | **Transform used (log or semilog)** | **Screenshot of transform plot** |
| Temperature | Semilog transform  LN(Y) vs X |  |
| Atemp | Semilog transform  LN(Y) vs X |  |
| Humidity | Semilog transform  LN(Y) vs X |  |
| Windspeed | Semilog transform  LN(Y) vs X |  |

Addressing Multicollinearity

1. Creating a correlation table for the independent variables.



1. Finding the independent variables that may be sources of multicollinearity.

|  |  |
| --- | --- |
| Temperature atemp | Temperature variable shows value close to 1 with atemp variable and are strongly positively correlated. They are a source of multicollinearity. |
| Season  Month | They might be a source of multicollinearity but we can disregard them because they are natural happenings. |

**Part Four—Validate Your Model**

|  |
| --- |
| UCI bike sharing dataset is used in this project to predict the bike rental demands. People tend to use this Bike Share either casual or with a paid membership. From registering to renting a bike and returning them back to a station is fully automatic. We have around 16 columns and 731 observations.  We will be eliminating Index, Casual, Registered, Year columns.  The index column doesn’t give valuable information. Dependent variable is our BikesCount variable because we are going to predict the bike rental demand.  As we have the Bikescount column which is the sum of Casual and Registered Column. We will eliminate them. Year variable ranges between 2011 and 2012. This could not be a reliable predictor of the dependent variable. We can use the Year column to check the popularity trend of BikeShare in the years 2011 and 2012.  With the scatterplot analysis, temperature and atemp variable shows positive association and have some effect on the dependent variable. At the same time windspeed and humidity shows a negative effect on them  With the regression analysis we can infer the following:  R-squared value being closer to 1, our model has some good variables for prediction and better regression equation. Significance f value is <0.05 also indicates that our model is ok  Coefficients of temperature and atemp variables are the highest among other independent variables. This also shows that they have a positive effect on the bike rentals demand. Coefficients of weather, holiday, windspeed, humidity shows negative effect on the bike rental demand.  Season, Weather, Holiday, Weekday, Workingday, Month are categorical variables. Temperature, Atemp, Windspeed, Humidity are continuous variables.  Temperature and atemp seemed to have multicollinearity. As atemp has values from Temperature variable, we will eliminate atemp variable. Humidity and windspeed shows a negative effect on the dependent variable. Windspeed has lower correlation as compared to temperature and humidity. But the correlation table indicated that Season and Month variable has strong correlation.  For the model we can consider the following independent variables. Temperature, Season, Working Day, Weekday, Humidity, Weather.  Running the regression test with the above-mentioned variables we have adjusted R-squared value and this model will help to explain around 50% of the variation of the dependent variable around the average values of the observations. Comparatively coefficients of the independent variables are larger than their standard error value. They are probably statistically significant.  From ANOVA test, p-value 3.1098E-108 of f-test is <0.05. Thus, we can reject the null hypothesis that all coefficients are equal to 0.  To conclude this, the number of bikes rented depends on weather and temperature values. Because they show high positive relationships. It is also mainly seen that average amount of bikes increases and decreases with the temperature. Thus, to predict bike rental demands and to start new stations in in new places these variables play a significant role. |