First Capstone Project Proposal

Project Idea 1: Predict the Bike Rental volume from the dataset given by Capital Bike Sharing System. Also determine the factors or features that influence Bike Rental Count most.

PROPOSAL DETAILS

1. What is the problem you want to solve?

I plan to analyze Bike Sharing system program data hosted by UCI
http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

This dataset contains the hourly count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information. The dataset contains nearly 17K records with 16 attributes.

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingda	weathersi	temp	atemp	hum	windspee	casual	registered	cnt
2	1	1/1/11	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0	3	13	16
3	2	1/1/11	1	0	1	1	0	6	0	1	0.22	0.2727	0.8	0	8	32	40
4	3	1/1/11	1	0	1	2	0	6	0	1	0.22	0.2727	0.8	0	5	27	32
5	4	1/1/11	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0	3	10	13
6	5	1/1/11	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0	0	1	1
7	6	1/1/11	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1
8	7	1/1/11	1	0	1	6	0	6	0	1	0.22	0.2727	0.8	0	2	0	2

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

The problem that I want to solve in this project is:

- Predict the Bike Rental volume from the dataset given by Capital Bike Sharing System. This dataset comprises two year worth of data from 2011-2012.
- Determine the factors or features that influence Bike Rental Count most.

2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

The client is Capital Bikeshare System and this research to predict bike rental count will be useful to them in knowing:

- What features in the dataset influence the bike rental count
- When is the demand for bike share program maximum during the day, season, quarter or year.
- Does weather conditions like temperature, humidity, windspeed have any impact on the demand? If yes, then is it to advantage or adverse.
- Are bike users whether Registered or Casual drive Bike Rental Count? If yes, do they have similar influence on Bike Rental Count distribution?

3. What data are you going to use for this? How will you acquire this data?

Capital Bikeshare posts quarterly data reports of bike trip times, start and end locations, and type of user (registered or casual). Each trip is on one line of data. These data are readily and publicly available at https://www.capitalbikeshare.com/system-data. The data look like this:

Duration	Start date	End date	Start station	Start station	End station r	End station	Bike number	Member type
367	4/1/18 0:00	4/1/18 0:06	31103	16th & Harva	31214	17th & Corcc	W20968	Member
653	4/1/18 0:00	4/1/18 0:11	31201	15th & P St N	31503	Florida Ave &	W20887	Member
598	4/1/18 0:02	4/1/18 0:12	31268	12th & U St N	31251	12th & L St N	W00133	Member
2015	4/1/18 0:03	4/1/18 0:37	31505	Eckington Pl	31228	8th & H St NV	W20534	Casual
563	4/1/18 0:04	4/1/18 0:14	31214	17th & Corcc	31223	Convention (	W23117	Member

4. In brief, outline your approach to solving this problem (knowing that this might change later).

The first goal is to exhaustively determine the various relationships among bike use, User Type and weather variables.

1. Data wrangling will occur to clean datasets to applicable variables and check for inconsistencies in Capital Bikeshare dataset as rename features, change feature data types as required.
2. Exploratory Data Analysis (EDA) will occur to check for possible trends and/or correlations between bike usage characteristics, User Type and weather variables. The basic preliminary questions should be confidently answered after this stage.

3. Statistical inferences to conclude ‘Rental Bike Count’ follow Normal Distribution curve thus indicating for any given period there is 95% Confidence interval of the value lying within the curve.
4. Generating multiple samples of mean Bike Rental Count at the Capital Bikeshare System. Remember, we are estimating the mean Bike Rental Count we would get if the Capital Bikeshare System could repeat all of the measurements from 2011 to 2012 over and over again. This is a probabilistic estimate of the mean & was proved through histogram showing Normal distribution curve. Bootstrap replicate function was used to run the measurements over & over & find mean for each sample.

5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

The final products will include

- code ipython notebook
- Final Report
- Presentation slides

6. Is this Supervised or unsupervised problem?

This is Supervised problem as the output datasets are provided and this I used to predict the future outcomes of target variable.

7. Is it a classification or regression problem?

This is regression problem as dependent variable i.e. Bike Rental Count is continuous values or ordered whole values. **Regression** means to predict the output value using training data.

8. What variable is it that you are trying to predict?

Bike rental volume is the target variable that I analyzed to find correlation, mean, standard deviation, minimum residual sum of squares to find optimal linear function for prediction with other related features in the dataset.

9. What variables will you use as predictors?

Following are independent variables that are influencing the outcome of target variable ‘ Bike Rental Volume’ in this project:

-Temperature

- Casual User Type

- Registered User Type

- Humidity

- Windspeed

- Hour of the day