Name: Robert Kaszubski

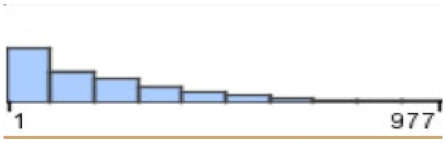
ID: 1866066

Bike Sharing - Project Report

My project was an analysis of the Bike Sharing Dataset found on the UCI Machine Learning Repository. What drew me to this dataset over some of the other one's I found was my love and prior knowledge about biking. As someone who bikes a lot, both for fun and for commuting, it was a topic I was immediately interested in exploring. Although I have rarely made use of a bike sharing program, I have seen many people make use of them here in Chicago. In Chicago we have our well-known Divvy bike sharing program, which is tied to the Chicago Department of Transportation. Chicago isn't the only city with a bike sharing program, in fact most major urban areas in America have them. This dataset is taken from Capital Bike Share, a bike sharing system servicing Washington, DC. The company posts their hourly data on their website for public use. That is where Hadi Fanaee-T from the University of Porto in Portugal, the publisher of this dataset on UCI, sourced the data from. The data set also consists of hourly weather data that was sourced from local Washington, DC news sites.

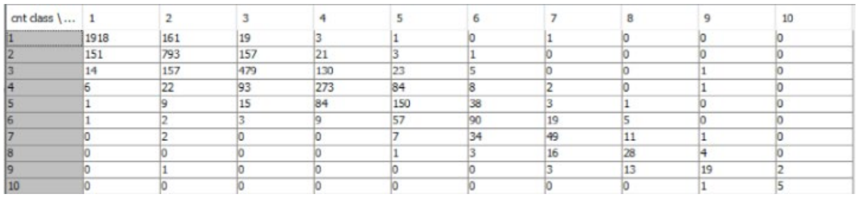
The data set consists of 17,389 instances, one for each hour of the days measured over the course of two full years. Each instance contains data under 17 different variables. The first variable is a string with the date, followed by the season, year, month, day of the week, hour. Next, we have two binary variables that indicate whether the day is a holiday or not, and whether the day is a working day or not. This is followed by all of the weather information which is broken down into 5 different variables: temperature, feels like temperature, humidity, windspeed (all four of which are already normalized using max min normalization), and a final class variable that categorizes whether the weather on the day was clear, misty and cloudy, had light rain, or had heavy rain + thunderstorms (extreme weather). All those variables would be explanatory variables and our outcome variables are our final three which show the total number of bike rentals (for casual riders, registered members, and both combined) for each hour. Looking at these variables, I decided that I wanted to explore the relationship between the weather and the time, and the resulting number of bike share rentals. This would provide a good understanding of the related factors that impact a person's inclination to make use of a bike share program. Besides this I wanted to look at the difference in casual and registered users, as well as how cycling habits change during the year. I think this data is particularly valuable from a business perspective as we are essentially analyzing the mobility in a city. Bike sharing businesses could gather regional data to determine which stations are the most used and could make better informed decisions on where to place new ones. It's also valuable in the explanation of human behavior and how we are affected by weather.

My work on this dataset was done entirely in Knime, a software designed specifically for this kind of data analysis. Having imported my dataset, I saw that for the most part the data set wouldn't require too much initial cleaning. There were no missing data points, so my only cleaning ended up being the removal of the date string variable. There was no use for it as it wasn't numerical data and wasn't useful as a category. From there, I generated a statistical summary. Looking at the outcome variable "count", the min was 1 and the max was 977 with an average of around 189. That was already interesting to me as no matter what time and no matter how bad the weather was there was at least 1 person renting a bike. But the more important part of the statistical summary was the histogram for the count variable which showed how skewed the data was. I



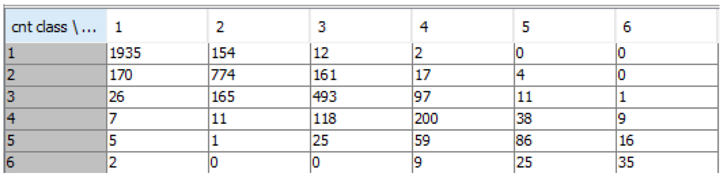
decided to create my own class variable called “count class” based on the histogram. To do so, I used a rule engine that categorized each instance into a class from 1-10. Each class is in increments of 100 from the count variable. Meaning that if the count was between 0-99 it was classified as 1, if it was 100-199 it was classified as 2, and so on. This new variable was created in order to be able to build a classification model.

Building the classification model took a lot of trial and error. The classification model was designed to predict the “count class” While it is easy enough to build in Knime, it was hard to determine the variables to include and exclude. My final model filtered out the count, casual, and registered columns and included everything else. The other columns were then all normalized if they weren’t already. From there the dataset was partitioned with about 70% going to the training phase and the other 30% going to the testing phase. I used stratified sampling in order to ensure that some of each class was used in the training phase. Had I used a strictly random sampling then it wasn’t guaranteed that the model would train using the higher classes of which there were less of. There were only a few instances that were classified as 9 and 10 in the initial dataset after all. From there it took a little tweaking to improve the accuracy. This involved making use of the force root split option which allowed you to set the first variable that the decision tree would split with. It was initially set to temperature, but my best result came from using the “weathersit” variable which was the variable that classified what the weather was like that day (clear, cloudy, rainy, severe). My model had an overall accuracy of 73% with the most important variables in the decision tree after the “weathersit” variable being hours and temperature. This was what I expected as people would be less likely to bike in the cold and at night. Looking at the confusion matrix



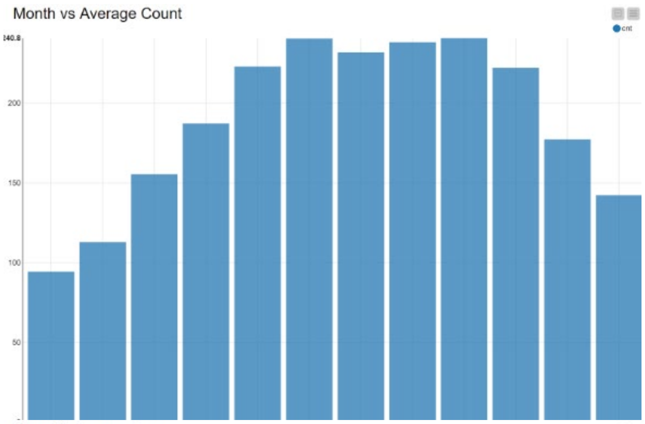
cnt class \...	1	2	3	4	5	6	7	8	9	10
1	1918	161	19	3	1	0	1	0	0	0
2	151	793	157	21	3	1	0	0	0	0
3	14	157	479	130	23	5	0	0	1	0
4	6	22	93	273	84	8	2	0	1	0
5	1	9	15	84	150	38	3	1	0	0
6	1	2	3	9	57	90	19	5	0	0
7	0	2	0	0	7	34	49	11	1	0
8	0	0	0	0	1	3	16	28	4	0
9	0	1	0	0	0	0	3	13	19	2
10	0	0	0	0	0	0	0	0	1	5

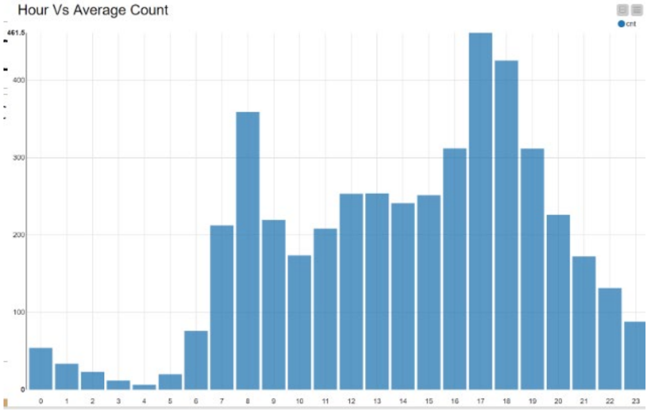
generated by my model and you can see that for the most part, the model accurately predicted which class each hour belonged in. The biggest surprise for me was my model accurately predicted 5 out of the 6 instances that were classified as 10, which I thought was impressive. Some of my earlier models were unable to do so and would generally classify everything between 1-5 and rarely predict anything higher than that correctly. You can also spot a few of the outliers for instance a 2 being predicted when it should have been a 9. This can mean that there were likely outside circumstances that led to the number being off by so much. I then decided to remove the outliers using 1.5 times IQR and rerun the same classification model. Outliers were typically caused by weather or factors inconsistent with what is expected during that time period. For example, that could mean a severe thunderstorm in the middle of a June afternoon when you expect bike share usage to be high but isn't due to the weather. Or on the contrary it could be a spike in users due to a special event in the city or other outside circumstances that encourage users to bike. This led to an improvement in accuracy to 75.5%, a minor improvement. As a result of the removal of outliers, the count class variable only had 6 classes rather than 10 as seen on the new confusion matrix. The important variables on the decision tree remained the same.

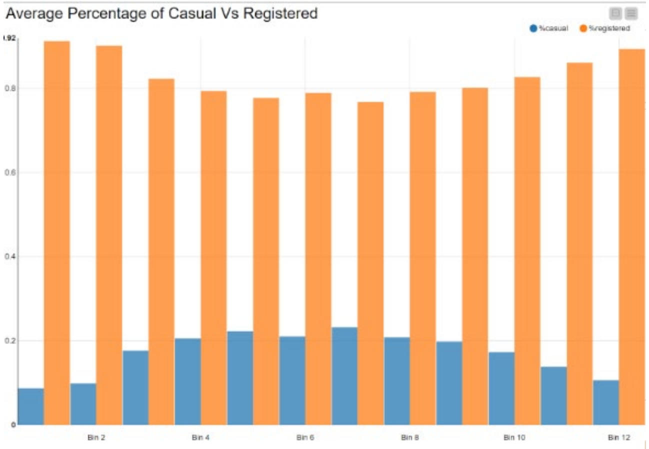


cnt class \ ...	1	2	3	4	5	6
1	1935	154	12	2	0	0
2	170	774	161	17	4	0
3	26	165	493	97	11	1
4	7	11	118	200	38	9
5	5	1	25	59	86	16
6	2	0	0	9	25	35

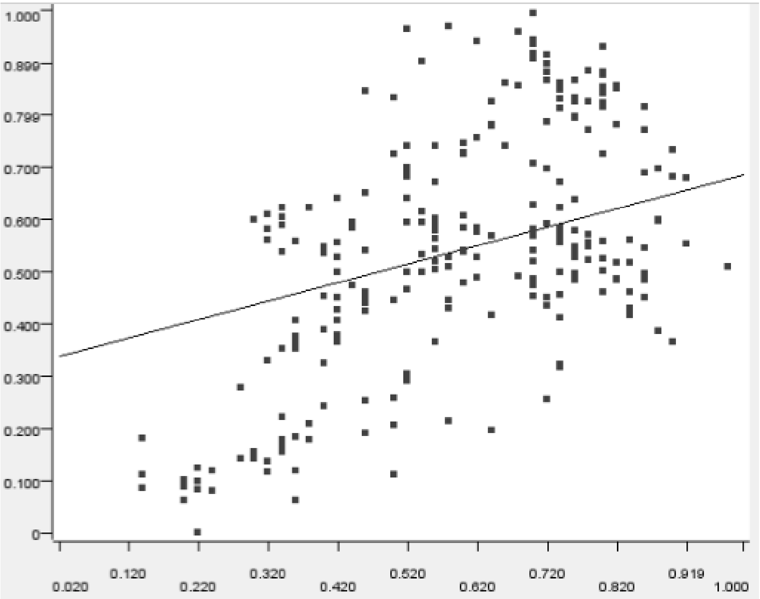
I also wanted to try using linear regression with my dataset to predict the count variable or the number of bike share rentals at an hour. I knew right away from my early test that it would be more difficult than classification due to the nature of my data. Having generated a few data visualizations, I saw how the count variable behaved in relation to hour and month. As you would expect, fewer people on average bike during the night, and peak hours are around 8 am and 5pm. This can easily be explained as the usual time people leave to work and leave from work, which makes sense when looking at the widely higher proportion of users being registered members rather than casual riders, meaning that bike sharing is used more for commuting than any other purpose. As a result, my data follows a wavelike near parabolic motion each day and over the course of the months (more people biking during the warmer months than the winter) which would make accurate regression impossible. This meant that I had to filter out a lot of data using the row filter nodes in Knime before creating my regression model. I filtered my data to only include working days, at the peak time of 5pm, and at an ideal weather situation (1 or



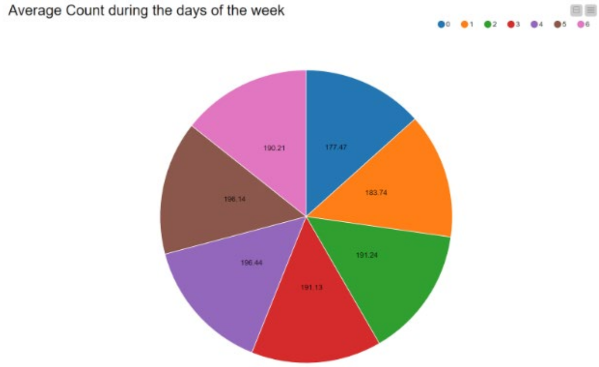


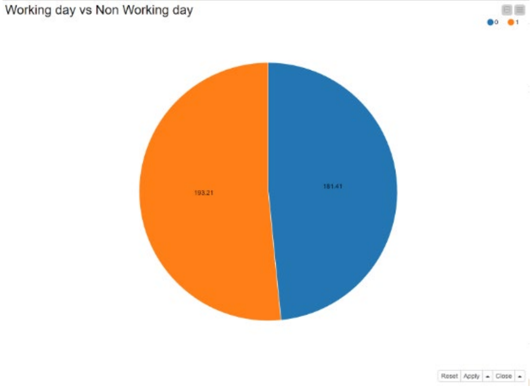


“clear”) to avoid any outliers. From there the data was normalized, partitioned the same way as the classification model and fed into a regression model. This also required a lot of trial and error before I ended up using only the weather variables meaning temperature, feels like temperature, wind speed, and humidity to predict the count. This gave the best results although still not ideal with an r value of around .56 indicating some correlation, meaning that weather does impact one’s desire to make use of bike sharing but isn’t necessarily a deal breaker. Think for instance of someone going biking when its 70 degrees versus 80 degrees Fahrenheit. It likely won’t change their decision but that difference in temperature does affect our regression model. Below is a scatter plot of my testing data set with the normalized temperature on the x axis and the count on the y axis.



Besides my two main models, I generated a few other data visualizations to better understand the data and answer my questions. Looking at the average users during the days of the week, I found the number to be about the same for each day. The least popular was Sunday (0 on the pie chart) at around 177 users while the rest of the week was all around 190. This surprised me as I had initially expected bike share usage to be higher on the weekends, but the data shows that people prefer to use it for commuting needs rather than casual riding for enjoyment. This is backed up by my other pie chart that looked at the average use on working days versus non-working days. Although the difference is minimal it's clear that on average more people using during working days.





My takeaways from all of this are that there is some correlation between temperature and bike share rentals, with weather and time of day being a strong indicator of how many people are making use of bike sharing. Most bike share users are registered members rather than casual riders with a strong implication that bike sharing is used primarily for commuting to work and school as evident by the peak hours of use. This is further shown with bike sharing used more often on working days than holidays, and a surprisingly equal amount of usage on each working day of the week.

Had I had more time I would have liked to have run an almost reverse classification model of the one I used here meaning that I would use the "weathersit" variable as my outcome variable to predict and use the time of day and number of bike share users (count) as part of my classification model. I would like to see if I could accurately predict the weather based on the number of people that are out biking. Additionally, I wish my dataset also had other weather variables such as rainfall which might have led to a stronger correlation. I think that would have been more useful than a simple weather situation variable which generalized if it was clear, cloudy or raining. In conclusion, I am pleased with how this project turned out and I think my findings prove what I set out to discover.

Literature Review

DeMaio, Paul. 2009. Bike-sharing: History, Impacts, Models of Provision, and Future.
Journal of Public Transportation, 12 (4): 41-56.
DOI: http://doi.org/10.5038/2375-0901.12.4.3
Available at: https://scholarcommons.usf.edu/jpt/vol12/iss4/3

One of my papers looks at the history of a bike-sharing system in Denmark over the past half century claiming that there have been 3 generations of bike sharing systems. The first of which was ordinary bikes painted all white that were provided for public use. It was a system were you found a bike, drove it to your destination, then left it for the next user. As you would expect, the system didn’t go well with many bikes being vandalized, stolen, and destroyed resulting in the system collapsing quickly. The second generation included designated areas to pick up and drop of each bike that were automated with a coin deposit. This still led to theft leading to improvements in customer tracking in bike sharing. The third generation resembles the bike sharing programs of today with bike sharing systems including trackers, electronic locking mechanisms and the use of a card or smart phone as payment.

Schuijbroek, J., Hampshire, R. C., & Hoeve, W.-J. van. (2016, August 17). Inventory rebalancing and vehicle routing in bike sharing systems. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0377221716306658?via=ihub.

This paper looks at the rise of bike sharing in popularity over the last few years. This has created challenges for bike sharing companies as they need to expand their infrastructure to support demand. This is like what I was looking into from a business perspective with how bike sharing programs can use datasets like mine in order to make business decisions in relation to where they should build new bike share stations. The paper explains that a big cost is indeed this kind of expansion and guaranteeing that the number of bikes available to consumers is adequate at each location and the paper explains potential solutions to this problem.

Vogel, P., Greiser, T., & Mattfeld, D. C. (2011, September 6). Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877042811014388.

This paper details the usage of data mining in understanding bike-sharing systems. It is essentially what I did in my project. They are looking at operational data from a bike sharing system to analyze patterns in usage. One of the big problems is the distribution of bikes, with certain stations having too many, and others not enough. Using data mining, they can look at the usage and plan ahead in that regard. This is essentially what I was proposing in my research project when discussing the application and value of my findings.

Midgley, Peter. (2011, May 2). Bicycle-Sharing Schemes: Enchancing Sustainable Mobility In Urban Areas. Retrieved from: https://www.un.org/esa/dsd/resources/res_pdfs/csd-19/Background-Paper8-P.Midgley-Bicycle.pdf

This final paper looks at the role that bike sharing plays in cities. This paper is looking specifically at European cities, but it applies to all parts of the world. Bike sharing is meant to be fast and easy access and now makes use of new technology while being integrate with other public transport systems. They are meant to introduce new mobility choices, as well as improve air quality. To implement bike sharing a city has to be committed to promoting cycling, implementing bike lanes and paths, and have enough area to construct bike sharing stations.

Reference List:

- DeMaio, Paul. 2009. Bike-sharing: History, Impacts, Models of Provision, and Future. Journal of Public Transportation, 12 (4): 41-56.
  DOI: http://doi.org/10.5038/2375-0901.12.4.3
  Available at: https://scholarcommons.usf.edu/jpt/vol12/iss4/3
- Midgley, Peter. (2011, May 2). Bicycle-Sharing Schemes: Enchancing Sustainable Mobility In Urban Areas. Retrieved from: https://www.un.org/esa/dsd/resources/res_pdfs/csd-19/Background-Paper8-P.Midgley-Bicycle.pdf
- Schuijbroek, J., Hampshire, R. C., & Hoeve, W.-J. van. (2016, August 17). Inventory rebalancing and vehicle routing in bike sharing systems. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0377221716306658?via=ihub.

- Vogel, P., Greiser, T., & Mattfeld, D. C. (2011, September 6). Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877042811014388.