Jay Filley, Michael Lunzer, Michael Yang: DSC 433

Eugene is ranked 6th in the nation for most bike thefts according to the Kryptonite Bike Lock Company. This is an amazing statistic considering that Eugene is the only city on that list that is not a heavily populated metropolitan area. When we took the class poll during our presentation, almost half of the class have had a bike stolen while living in Eugene, so this is a very prevalent issue in the community. We were even surprised by the significant number of bike thefts.

Why is Eugene so high on this list and how can we stop the rapid increase of bike theft? Many of these stolen bikes are sold in other cities or scrap mettle for what is becoming a very profitable business at the expense of the Eugene bike owners. Last year in Eugene, over 830 bikes were stolen. And of those 830 only 4% or 33 bikes were found and returned to their owners.

Initially our group was not sure which direction we wanted to go with our project, but we knew that we wanted to do something crime related in Eugene. We were fortunate enough to meet with Tim Schuck, the Crime Analysis Unit Supervisor over at the Eugene Police Department. He informed us about the huge bike theft problem all over Eugene and was wondering if we could use our data mining knowledge to help. The Eugene Police Department just did not have enough time to look at and decipher all of this data to predict and track bike thefts. This week, we are presenting our findings to the Eugene Police Department to offer them some actionable solutions.

Tim reiterated the point that jail space is a huge problem in Eugene. So, if the police catch a bike thief, he/she is usually out on the streets again in a few hours. If we can curtail the thefts and stop them from happening in the first place, the lack of jail space will not be a problem.

Tim gave us the bike theft data in Eugene from 2007-2012 so we could extract patterns and make predictions about future bike thefts. The tools that we used to accomplish this were data cleaning, multiple linear regression, data visualization, and we created a scatter plot map to get familiar with the data.

Data Exploration:

We went to the Eugene Police Department to meet with Crime Analysis Unit Supervisor Tim Schuck. He provided us with data for bike theft in Eugene over the past 5 years (2007-2012).

In order to better understand the data, some simple definitions of terms need to be explained.

With theft, there are 3 levels:

- Theft I is greater than $1000 dollars
- Theft II is less than $1000 dollars, but greater than $100
- Theft III is less than $100 dollars

From this knowledge, we can create a rough estimate of the cost of the bikes the thieves are targeting. These ranges are large, however the prices are just an estimate from the owner.

Split time: it is the median time between when the victim last saw their stolen bike and when they noticed it missing.

- **Example**: Sheila locks up her bike outside her dorm after returning from the library at 9 PM. The next day when she goes to class at 9 AM, she goes to her bike and notices it is gone. She calls the police to file a report. The split time would be 2 AM.
    - Now this does not necessarily mean that the crime happened at this exact time, but it gives a good estimate and the police know it happened when it was dark and probably when most people were in bed.
- If the split time is over 24 hours, the EPD discards this value because it is extremely inaccurate and therefore useless.

Cleaning the Data:

When our group first received the data from the Eugene Police Department, it needed to be cleaned. Initially, the date and time were in the same cell. We separated the values using the “Text to Columns” feature in the Data tab of Excel. This allowed us to create a column for the date and time. From here, this allowed us to create columns for the day of week, represented with the numbers 1 through 7 starting with Monday. We also created columns for Month, Day, and Year.

Relevant Variables:

To gain a clearer picture of our information, we needed to break our data down into just the relevant data to what we were trying to explore. We trimmed out the variables which were irrelevant for that specific analysis that we were running.

Some examples of useless variables include:

· **Incident Address & Zip Code** – we opted to use the more accurate X and Y GPS coordinates when plotting the bike thefts.

· **Premise** – This provides a description of the area (parking lot / sidewalk / Apartment Complex / College).

Then we determined many useful variables that helped us in our model formation. Some examples of this include:

- **Geographic Areas** – shown as X & Y coordinates to plug into a map.
    - We used this data to map the points on top of a map of Eugene.
- **Beat Areas** – Eugene is split into 6 “beats” of varying populations and sizes.
- **Split Time** – This time from bike last seen to realized stolen.
- **Charge Type** – which gives us a range of dollar bike value.

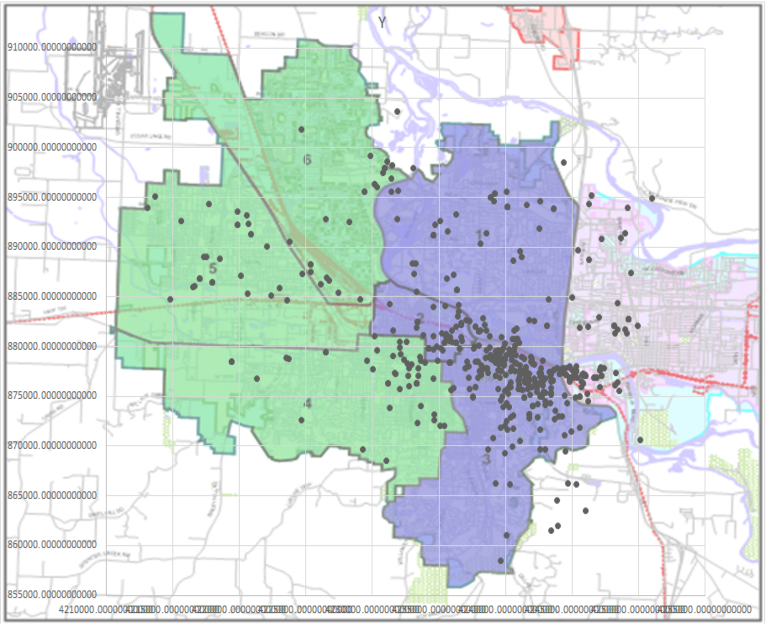
- We created dummy variables for the Day of Week, Month, Beat Area, Holiday, and Home Football Game.

Summarizing the Data:

Once we had cleaned the data, we created a summary table that counted the number of bike thefts on each day of the year over the five-year span and a breakdown of how many bike thefts occurred in each beat. Also, we created columns with the day of week, represented by a 1 through 7, starting with Monday. Additionally, we created columns to see if that day was a national holiday or a home football game, to see if any correlations can be drawn from those instances. These columns contained a binary response, so 0 if it was not a holiday or home football game and 1 if it was.

After this, we used the X & Y coordinates and plotted them on top of the beat map of Eugene. This allowed us to see where different clusters lie on the map.

We determined using Multiple Linear Regression the day of week and locations which have the highest occurrences of bike theft. Our results were visually confirmed on the scatter plot map shown below.

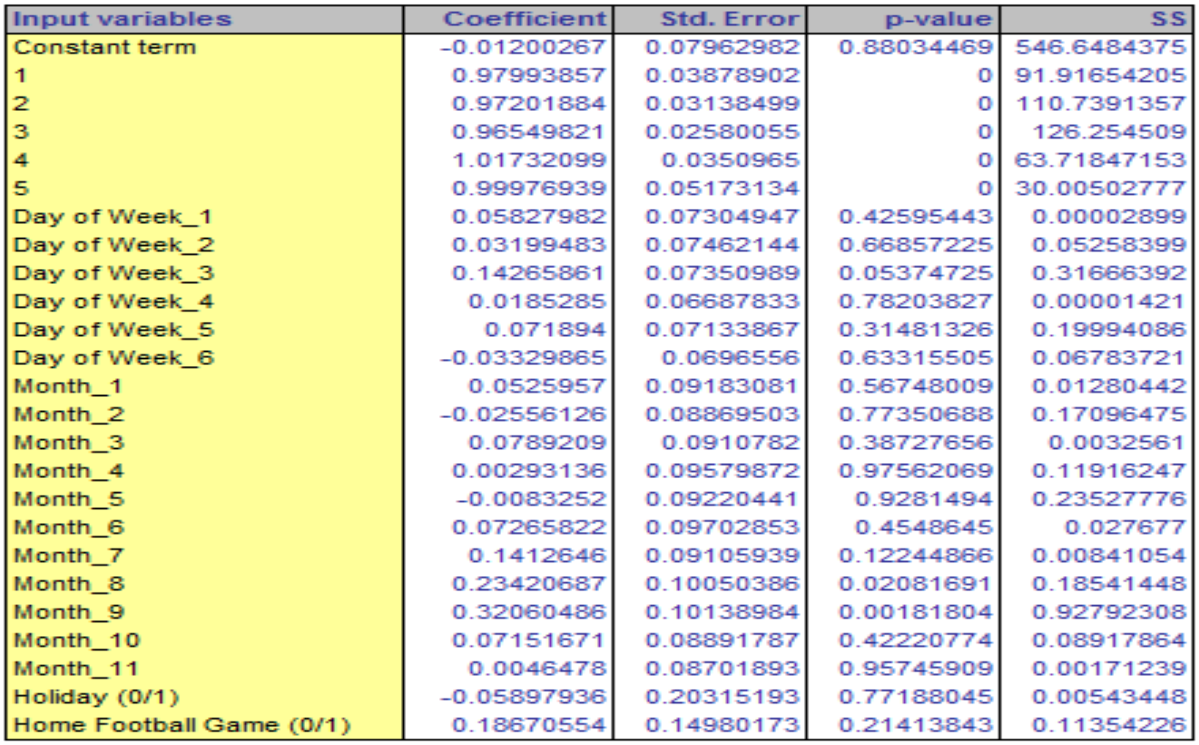


It is clear that the highest areas of bike theft occur in Beat 3, which is the campus area of Eugene. This makes sense because there is a huge population of bike riders and therefore opportunities for bike thieves.

Model Development, Estimation, Interpretation

The model we choose to use is multiple linear regressions (MLR). The reason why we are using MLR is because we want to predict where and when a bike theft will occur. We wanted to try other models with our data set but it wasn't possible. Other models we tried using were logistics regression and classification/regression trees (CART). We thought we could use multiple models to have a similar outcome as our question but we were wrong. After running our data through these models we had high p-values and our outputs did not make any sense. After using these models we decided to stick with MLR because it helped answer our question better than other models. The variables we used were: days of the weeks, months, beats, home football games, and holidays.

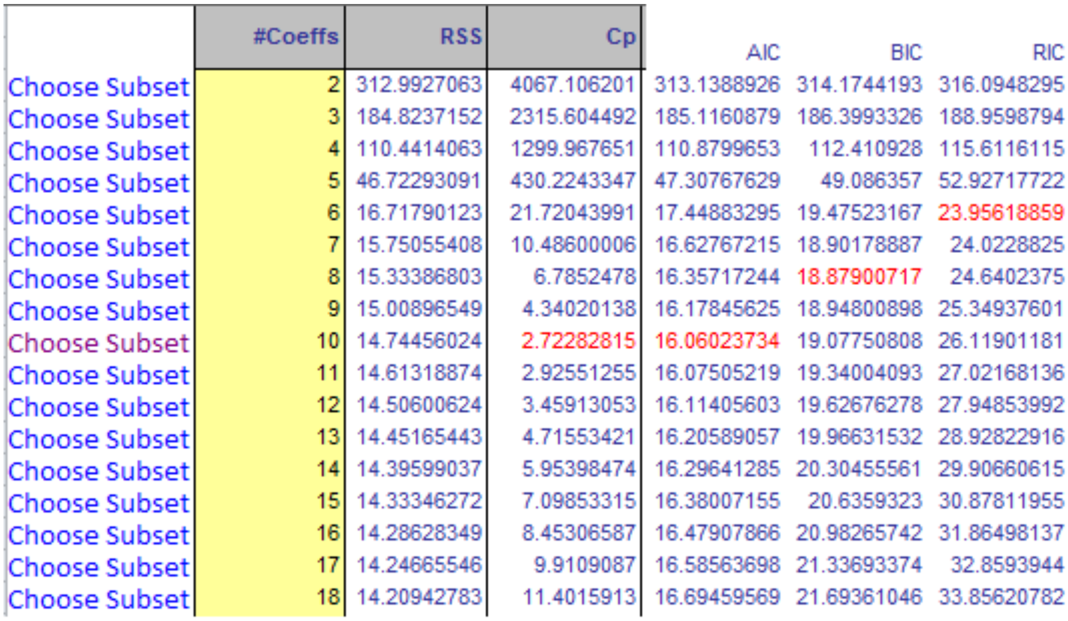
As explained previously before we ran our data through our models we had to clean, sort, and sum our data. After that we needed to create dummy variables so days of the week and months would represent Monday through Sunday and January through December. Next we partition our data and ran our data through the MLR with best subset and a detail report. Selecting best subset would allow us to eliminate useless variables and a detail report would help us find which model would have the lowest error rate.



Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-0.01200267	0.07962982	0.88034469	546.6484375
1	0.97993857	0.03878902	0	91.91654205
2	0.97201884	0.03138499	0	110.7391357
3	0.96549821	0.02580055	0	126.254509
4	1.01732099	0.0350965	0	63.71847153
5	0.99976939	0.05173134	0	30.00502777
Day of Week_1	0.05827982	0.07304947	0.42595443	0.00002899
Day of Week_2	0.03199483	0.07462144	0.66857225	0.05258399
Day of Week_3	0.14265861	0.07350989	0.05374725	0.31666392
Day of Week_4	0.0185285	0.06687833	0.78203827	0.00001421
Day of Week_5	0.071894	0.07133867	0.31481326	0.19994086
Day of Week_6	-0.03329865	0.0696556	0.63315505	0.06783721
Month_1	0.0525957	0.09183081	0.56748009	0.01280442
Month_2	-0.02556126	0.08869503	0.77350688	0.17096475
Month_3	0.0789209	0.0910782	0.38727656	0.0032561
Month_4	0.00293136	0.09579872	0.97562069	0.11916247
Month_5	-0.0083252	0.09220441	0.9281494	0.23527776
Month_6	0.07265822	0.09702853	0.4548645	0.027677
Month_7	0.1412646	0.09105939	0.12244866	0.00841054
Month_8	0.23420687	0.10050386	0.02081691	0.18541448
Month_9	0.32060486	0.10138984	0.00181804	0.92792308
Month_10	0.07151671	0.08891787	0.42220774	0.08917864
Month_11	0.0046478	0.08701893	0.95745909	0.00171239
Holiday (0/1)	-0.05897936	0.20315193	0.77188045	0.00543448
Home Football Game (0/1)	0.18670554	0.14980173	0.21413843	0.11354226

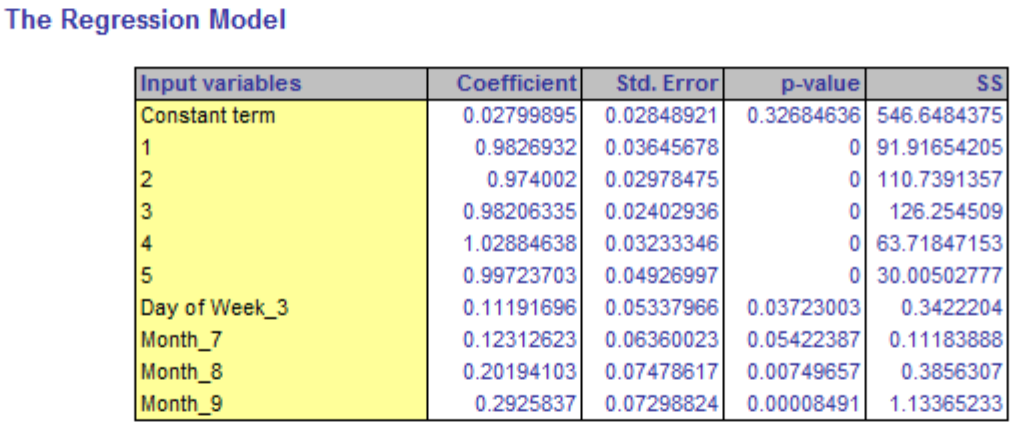
Full Model Equation-# of Crimes(x)=Constant Term+1(x)+2(x)+3(x)+4(x)+5(x)+Day of week_1(x)+…..+week_6(x)+month_1(x)+…..+month_11(x)+Holiday(0/1)+Home Football Game (0/1)

With the model above we are able to predict where a bike theft will occur in each beat, the day of the week and which month it'll occur, also if more thefts will occur during home football games and holidays. If we used the full model and not eliminate useless variables there will be problems. Over fitting occurs if the model is too complex or there are too many parameters. If we use the full model the model will have poor predictive performance. Looking at the full model there are variables with a p-value that is close to 1. Having a p-value that is close to or over 1 means that it is a useless variable and should be taken out of the model. For example the variable Month_4 has a p-value of 0.97. Having useless variables left in our model would just hinder our predictive performance.



	#Coeffs	RSS	Cp	AIC	BIC	RIC
Choose Subset	2	312.9927063	4067.106201	313.1388926	314.1744193	316.0948295
Choose Subset	3	184.8237152	2315.604492	185.1160879	186.3993326	188.9598794
Choose Subset	4	110.4414063	1299.967651	110.8799653	112.410928	115.6116115
Choose Subset	5	46.72293091	430.2243347	47.30767629	49.086357	52.92717722
Choose Subset	6	16.71790123	21.72043991	17.44883295	19.47523167	23.95618859
Choose Subset	7	15.75055408	10.48600006	16.62767215	18.90178887	24.0228825
Choose Subset	8	15.33386803	6.7852478	16.35717244	18.87900717	24.6402375
Choose Subset	9	15.00896549	4.34020138	16.17845625	18.94800898	25.34937601
Choose Subset	10	14.74456024	2.72282815	16.06023734	19.07750808	26.11901181
Choose Subset	11	14.61318874	2.92551255	16.07505219	19.34004093	27.02168136
Choose Subset	12	14.50600624	3.45913053	16.11405603	19.62676278	27.94853992
Choose Subset	13	14.45165443	4.71553421	16.20589057	19.96631532	28.92822916
Choose Subset	14	14.39599037	5.95398474	16.29641285	20.30455561	29.90660615
Choose Subset	15	14.33346272	7.09853315	16.38007155	20.6359323	30.87811955
Choose Subset	16	14.28628349	8.45306587	16.47907866	20.98265742	31.86498137
Choose Subset	17	14.24665546	9.9109087	16.58563698	21.33693374	32.8593944
Choose Subset	18	14.20942783	11.4015913	16.69459569	21.69361046	33.85620782

By choosing the best model using Mallow's Cp, AIC, BIC, RIC, we can eliminate useless variables. For Mallow's Cp we simply find the smallest number in the best subset table to find the best model chosen by Mallow's Cp. With AIC, BIC, and RIC we had to use a formula to find the smallest number and see which model would be best. The best model chosen by Mallow's Cp and AIC is the model with 10 coefficients.



The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	0.02799895	0.02848921	0.32684636	546.6484375
1	0.9826932	0.03645678	0	91.91654205
2	0.974002	0.02978475	0	110.7391357
3	0.98206335	0.02402936	0	126.254509
4	1.02884638	0.03233346	0	63.71847153
5	0.99723703	0.04926997	0	30.00502777
Day of Week_3	0.11191696	0.05337966	0.03723003	0.3422204
Month_7	0.12312623	0.06360023	0.05422387	0.11183888
Month_8	0.20194103	0.07478617	0.00749657	0.3856307
Month_9	0.2925837	0.07298824	0.00008491	1.13365233

The reason we didn't use BIC and RIC because it was a little too aggressive. We didn't want to eliminate too many variables. For example if we decided to use BIC and RIC to choose our model we wouldn't have Day of Week_3 in our model, and we would like to include what day a bike theft would occur.

Refine Model by Mallow's Cp, AIC-# of crimes(x) =Constant term+1(x)+….+5(x)+Day of Week_3(x)+Month_7(x)+Month_8(x)+Month_9(x)

The refine model does not have either holiday or a home football game. However the model including home football game has a Cp of 2.9255 vs. 2.7228. Also AIC doesn't differ much between the two models. The AIC model with home football game is 16.0750 vs. 16.0602. The different is roughly .015. So if needed, predicting bike thefts during a football game using this model would work also.

As mention before our model can predict how many bikes are stolen in any given day, month, and beat. For example on a Wednesday in the month of September on UO campus how many bikes are stolen in a day?

Using our model - # of crimes(x) =Constant term+1(x)+….+5(x)+Day of Week_3(x) +Month_7(x)+Month_8(x)+Month_9(x)

# of crimes(x) =0.02799895+.98206335+0.11191696+0.2925837=1.41

According to our model on a Wednesday in the month of September on UO campus roughly 1.41 bikes will be stolen in a given day.

Model Validation

Full Model-MAPE=0.08864 RMSE=0.26931

Best Model- MAPE=0.07938 RMSE=0.25948

As shown above the best model has the lowest MAPE and RMSE. Meaning it'll be the preferred model to use when finding when bike thefts will occur. Something to know is that the full model has a relatively similar MAPE and RMSE compared to the best model. They only differ by 0.00926 for MAPE and 0.00983 for RMSE. There isn't all that much of a different between the full model and best model in terms of error rate.

Conclusion

Based on these models that we created, we think that we have come up with actionable data and suggestions for the police department to use. The data mining tools that we used helped us wean out important data to better understand it as a whole.

We initially struggled with the data because there was so much of it and we could not come up with any actionable suggestions or solutions. However, after analyzing the data, we concluded that there are patterns to the bike thefts around the city and that some thefts can be prevented if there is some indication as to when and where they happen. If the police use this data to form their patrol patterns, many of these thefts will not happen and the Eugene bike theft problem will decrease dramatically.

Prevention is key when trying to reduce bike theft in Eugene. Due to the overcrowding in local jails, most criminals are released the same day because bike theft does not carry the same weight as for example, a criminal who hurts people. If we can create a culture of prevention among the police and bike riders, then the criminals will become discouraged when opportunities diminish. Eventually, this will reduce bike theft in the most concentrated areas and make Eugene a safer place to ride a bike.