SAS® GLOBAL FORUM 2018

USERS PROGRAM

Analyzing theft occurrences in Chicago

Shikha Prasad (Oklahoma State University),

Kunal Parekh (Verisk Analytics)

April 8 – 11 | Denver, CO

#SASGF

Analyzing Theft Occurrences in Chicago Using SAS® Enterprise Miner™ and SAS® Enterprise Guide®

Shikha Prasad, Kunal Parekh

Oklahoma State University, Stillwater, Oklahoma, USA

ABSTRACT

- In 2016, Chicago reported **more thefts than any other crime**, making it one of the most prevalent types of crime in the city. This poster explores patterns related to thefts in Chicago that can help minimize theft occurrences by answering questions such as: **What are the specific locations in the city where most thefts are committed? What could be the possible reasons for the frequency of thefts being higher in those locations than other crimes?** Crime data for the year 2012 through 2016 was obtained from kaggle.com [1], a publicly available data source, which had more than a million observations. Several predictive models such as logistic regression, decision tree, neural network and ensemble models were built to predict the target. For creating the binary target**, FBI code 06 (larceny) and 07 (motor vehicle theft) were assigned a value of 1 while all other crimes were assigned a value of 0.** The models were compared using the model comparison algorithm and the ensemble model emerged as the best model with the largest ROC index at 0.822. A **time-series** data was also prepared by aggregating the total number of theft incidents for each month to **forecast the number of thefts likely to take place in the next 12 months**. The stepwise autoregressive model generates the best forecast with a MAPE of 7%.

METHODS

Data Preparation and Analysis

- Data provided by the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting (CLEAR) system originally had 23 variables and 1,456,714 observations
- In Base SAS 9.4, a new variable was created to obtain the month from the date variable which had date time stamp
- To predict the outcome of a crime being a Theft or not, several models were built using SAS Enterprise Miner 14.1
- The data mining database (DMDB) node was used to study factor-level information and summary statistics for nominal and interval variables in the data
- Missing values existed for the variables – Latitude, Longitude, Location, Community Area Number, Ward, etc. For the analysis purpose, 40 missing values for Community Area Number were replaced by a value of “99” using a replacement node. Similar replacements were done for Ward, which had 14 missing values and District, which had 1 missing value. This was done so that in the regression model, observations with the missing values are not rejected as incomplete cases
- Interval variables, with a large percent of missing values such as latitude, longitude, X and Y coordinates were rejected from the analysis and imputation or replacements were not performed for these variables
- Data was partitioned into 60% train and 40% validation for obtaining proper model assessment. In order to predict the response, three models were built using the Regression, Decision Tree and Neural Network nodes. An ensemble model was also built to combine predictions from these models. The Model Selection node was used to select the model with the largest value of ROC index

METHODS CONTINUED

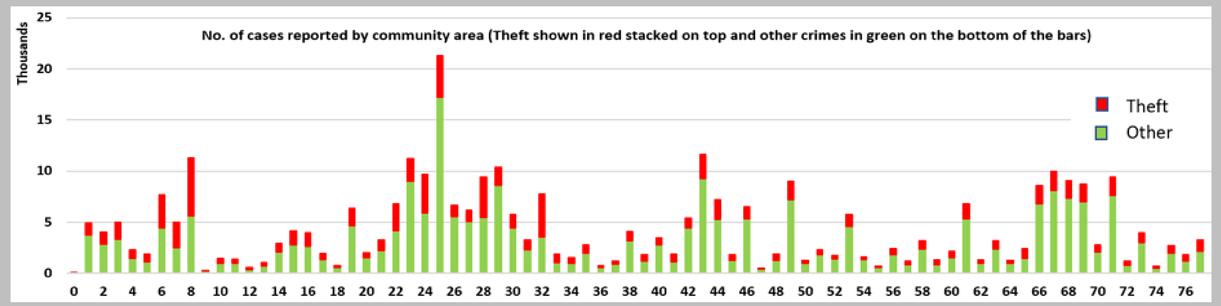
- Number of crimes were aggregated monthly for each year and forecasting algorithms were run in SAS EG 7.1 – **Analyze → Time Series → Basic Forecasting**

```
proc sql;
create table sp.ts_analysis_theft
as
select count(*) as N, year, mon,
date_new
from sp.analysis1
where theft = 1
group by mon, year;
quit;
```

```
data sp.ts_analysis2_theft;
set sp.ts_analysis_theft;
ts = mon || year;
run;
proc sort data = sp.ts_analysis2_theft
out = sp.ts_analysis3_theft nodupkey;
by ts;
format date_new monyy.;
run;
```

RESULTS

- The proportion of thefts out of all crimes is highest in community areas – 32 (Loop), 7 (Lincoln Park) and 8 (Near North Side)
- Most of the thefts are reported in the month of **July and August**
- Thefts are most common in **athletic clubs and department stores**
- Most of the thefts (89.5% of all thefts) do not result in arrests. **Only approximately 10% of all thefts resulted in some form of legal detention**
- From the results of the regression model – variables that are significant in predicting whether a crime will be identified as theft or non-theft are **Arrest, Domestic, Location Description, Community Area Number and Month**

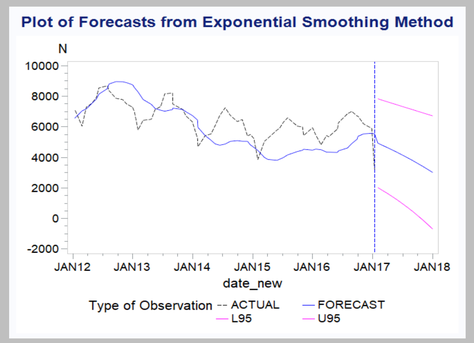


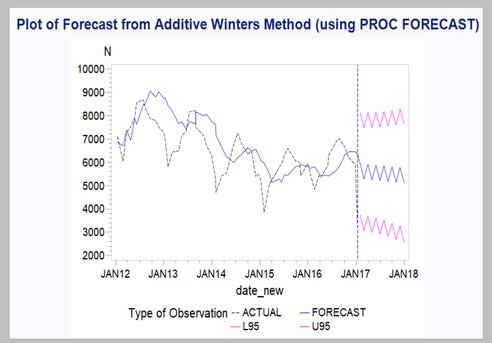
Analyzing Theft Occurrences in Chicago Using SAS® Enterprise Miner™ and SAS® Enterprise Guide®

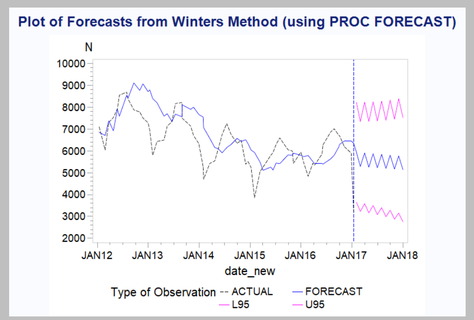
Shikha Prasad. Kunal Parekh

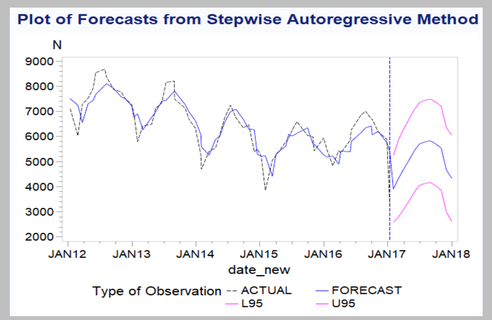
Oklahoma State University, Stillwater, Oklahoma, USA

FORECASTING RESULTS









CONCLUSIONS

- Community areas that are likely to report higher occurrences of thefts in Chicago are Lincoln Park (Community Area #7), Near North Side (Community Area #8) and Loop (Community Area #32). These are located in the North and Central regions of the city. The interesting fact here is that these three areas can be traced on the map, just one below the other. **Effective policing in these areas can help in minimizing theft occurrences**
- Using census data obtained from Chicago Data Portal that contains six socioeconomic indicators of public health significance and a hardship index by community area [2], it can be concluded that the areas where thefts are most prevalent have low value of hardship index reported. This means that **thefts occur primarily in those areas which are more developed and where the per capita income is greater than $65,000**
- Only about **2.5% of all the thefts are reported as Domestic**, which means that most of the theft occurrences are not related to households but instead take place outside homes in locations such as street, sidewalk and parking lot
- **Thefts in general, result in lesser number of arrests made as compared to other crimes**. This can be one of the possible reasons why the no. of thefts outnumber other crimes in the city
- Specific places where thefts are more frequently reported include **athletic clubs, department stores, food and drug stores, airport terminals, delivery truck and commercial vehicle locations and residential driveways**.
- The forecasts show the existence of a **seasonal pattern** as well as a **slightly decreasing trend**. The number of thefts starts dipping from the month of September and continues to decline until February, after which it rises again till August. Among all the four methods, **stepwise autoregressive** has least values for the residuals and mean absolute percent error (MAPE).

FORECASTING METHOD	MAPE
Exponential Smoothing	18.07
Winter's Additive	14.38
Winter's Multiplicative	14.20
Stepwise Autoregressive	7.40

REFERENCES

[1] Kaggle.com. (2017). *Crimes in Chicago | Kaggle*.: https://www.kaggle.com/currie32/crimes-in-chicago

[2] Data.cityofchicago.org. *Census Data - Selected socioeconomic indicators in Chicago, 2008 – 2012 | City of Chicago | Data Portal*.: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

ACKNOWLEDGEMENT

Thank you, Dr. Goutam Chakraborty (Director, Business Analytics, Oklahoma State University) and Dr.Miriam McGaugh (Clinical Professor, Business Analytics, Oklahoma State University) for your constant guidance and support throughout the research

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF