Analysis, Results and Findings:

There are many types of crime, so I divided them into two groups, which are petty crime and intense crime. And intense crimes are ARSON, ASSAULT, BATTERY, BURGLARY, CRIM SEXUAL ASSAULT, CRIMINAL DAMAGE, CRIMINAL TRESPASS, DECEPTIVE PRACTICE, INTERFERENCE WITH P, INTIMIDATION, KIDNAPPING, MOTOR VEHICLE THEFT, NARCOTICS, OBSCENITY, OFFENSE INVOLVING C, PROSTITUTION, ROBBERY, SEX OFFENSE, THEFT, and WEAPONS VIOLATION. The rest of crimes are petty crimes which are CONCEALED CARRY LIC, GAMBLING, HOMICIDE, LIQUOR LAW VIOLATIO, OTHER OFFENSE, PUBLIC PEACE VIOLAT, STALKING. And I created dummy variable for crime type as a binary variable.

There are over 6 million records in the data, and I selected a random sample in 2000 records size as my analysis dataset. My potential independent variables are Date, Community Area, Domestic, and Arrest. I created a dummy variable for Date as weather, which means when month = November, December, January, February, March, April and May, we see weather as Cold Weather. The rest months are Warm Weather. I also created dummy variables for community area. I divided 77 communities into 4 groups which are North, South, Northwest and Southwest based on community locations in Chicago. I created dummy variables for Domestic and Arrest as well to be part of my independent variables.

Because my dependent variable is a binary variable, I used logistic regression and there is no need to check the descriptive which make no sense. I just started with checking collinearity and outliers and collinearity looked good but there are many outliers, so I removed them before next step.

I split dataset into two subsets which are training set and test set with 75% sample rate. I fitted the model using training set with stepwise and forward methods for model selection and I got same results of remaining variables. The rest variables have significant parameters according to the p-value and the likelihood ratio in stepwise is larger compared to forward method with p-value almost zero. So, we can reject the null hypothesis that all parameters in the model are equal to zero. The AIC and SC also looked good. Therefore, the model is good for prediction. Then I checked collinearity, outliers and residual plots and they all looked good. Therefore, I decided my final model in this step (Appendix I, J, K).

I computed predicted probability for test set by finding the ideal cut-off value which is 0.94 (Appendix L) to classify the probability of Y into either 1 or 0. Then I compared observed Y with Predicted Y. They looked identical so the model is good. Finally, I computed the classification matrix to see the performance. Sensitivity or recall = 0.6, Classification Accuracy = 0.61, Precision = 0.986 and F-measure is 0.75 (Appendix M). I could conclude that Proportion of correctly classified positives among all positives and proportion of correctly classified positives and negatives among all cases looked just okay however the proportion of true positives among all predicted positives looked pretty good so overall the model's performance is okay.

Final model:
*Log (p/(1 - p)) = 2.5315 + 1.8221*Community_Area_N + 2.3986* Community_Area_SW + 1.0201*d_Domestic – 0.834*Weather + e*

Where:
Community_Area_N = 1(If crime happened in the north part of chicago) or 0 (Not in north part)
Community_Area_SW = 1(If crime happened in the southwest part of chicago) or 0 (Not in southwest part)

d_Domestic = 1 (The incident was not domestic-related) or 0 (was domestic-related)
Weather = 1 (if incident happened in months of 11, 12, 1, 2, 3, 4, 5) or 0 (incident happened in months of 6, 7, 8, 9, 10)

95% confidence intervals for Weather:
Weather = 1
The odds of crime being intense decrease about 56.6% between 23.3% and 75.4%.