karisteph /
Chicago_Car_Crashes_Analysis

<> Code | ⊙ Issues | ⑂ Pull requests | ▶ Actions | ⊞ Projects | 📖 Wiki | ⊘ Security | 📈 Insights | ⚙ Settings

👁 | ⑂ | ☆

☆ **0** stars   ⑂ **0** forks   ⊙ **1** watching   ⑂ **1** Branch   🏷 **0** Tags   〰 Activity

🌐 **Public repository**

⑂ main ▾ | ⑂ **1 Branch** | 🏷 **0 Tags** | ⑂ | 🏷 | 🔍 Go to file [t] | Go to file | + | Add file ▾ | Code | ⋯

🟣 **karisteph** Update README.md		2 minutes ago ⚫⚫⚫	🕐
📁 Data	uploading excel file with images for pr...	yesterday	
📄 .gitignore	Adding notebook and Data folder. CSV...	3 days ago	
📄 Chicago_Car_Crashes_Analysis...	uploading final notebook executed an...	3 hours ago	
📄 Chicago_Car_Crashes_Analysis...	uploading final notebook executed an...	3 hours ago	
📄 Chicago_Car_Crashes_Analysis...	Add files via upload	10 minutes ago	
📄 Chicago_Car_Crashes_Analysis...	Add files via upload	11 minutes ago	
📄 README.md	Update README.md	2 minutes ago	

📖 **README**   ✏ ☰

Chicago Car Crashes Analysis

Introduction

It comes as no surprise that the largest contributors to car accidents are neglecting safety measures and driving while incapacitated. Night-time driving and crashes in disadvantaged areas also significantly increase accident risks. Night driving raises fatality or serious injury chances by 90%, while disadvantaged areas add another 28%!

This project analyzes reported vehicle accidents in Chicago from the previous 9 years. We processed data from the city's data portal to provide insights for the city's current mayor.

Business Problem

Mayor Brandon Johnson is interested in reducing the deadliness of car crashes in Chicago. Our team was tasked with reviewing and analyzing the data from 2015 to the present to enhance road safety.

The mayor wants to know:

- Which factors have the greatest influence on fatal or serious injury car accidents?
- Are there any factors unrelated to driving behavior that contribute to serious car accidents?
- What measures can be implemented to decrease the likelihood of serious crashes occurring in the first place?
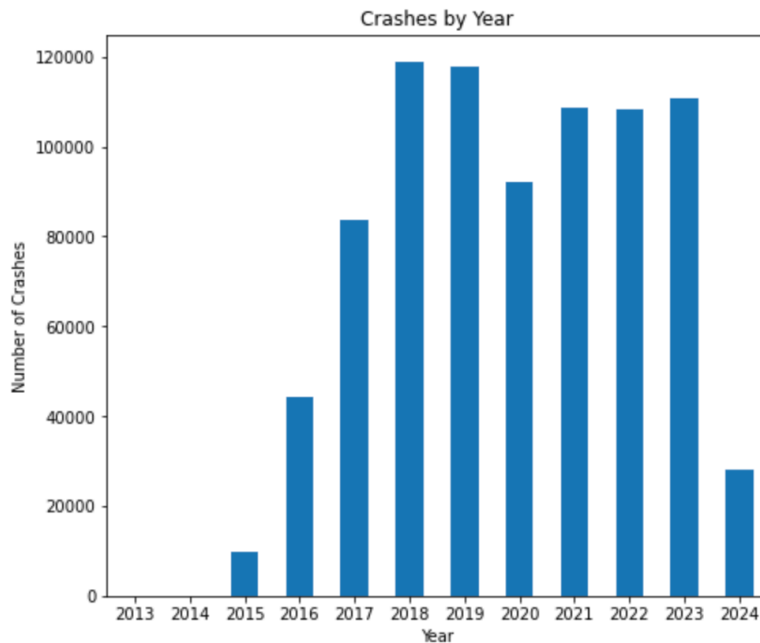
Data and Resources Used

For this project, we obtained a dataset from the Chicago Data Portal, encompassing data from 2015 up to April 2024. It includes:

- Car crashes
- People in crashes
- Socially disadvantaged areas geographies

Methods

We address our business problem by employing data science techniques, beginning with data cleaning.

- The initial data exploration included: 48 columns in the Crashes dataset and 29 columns in the People dataset
- Most columns were not useful
- Other columns have substantial gaps in the data
- We decided to keep 13 columns from Crashes and 3 columns from People

After cleaning the data we generated the master dataset and conducted an analysis using Python.

Further exploration of the data led to the application of more advanced modeling techniques to determine the likelihood of a crash resulting in a fatality or incapacitating injury.

The target variable is severely imbalanced, with only 1.8% of accidents involving severe consequences. As a result, accuracy of any model is likely to be naturally high, and also a bad metric. Instead, recall more closely fits the City's goals. At the expense of some false predictions, models targeting an improved recall score will provide actionable insights to the City to minimize these types of accidents.

Socially Disadvantaged Districts Map: the parts of Chicago that have been deemed 'Socially Disadvantaged'
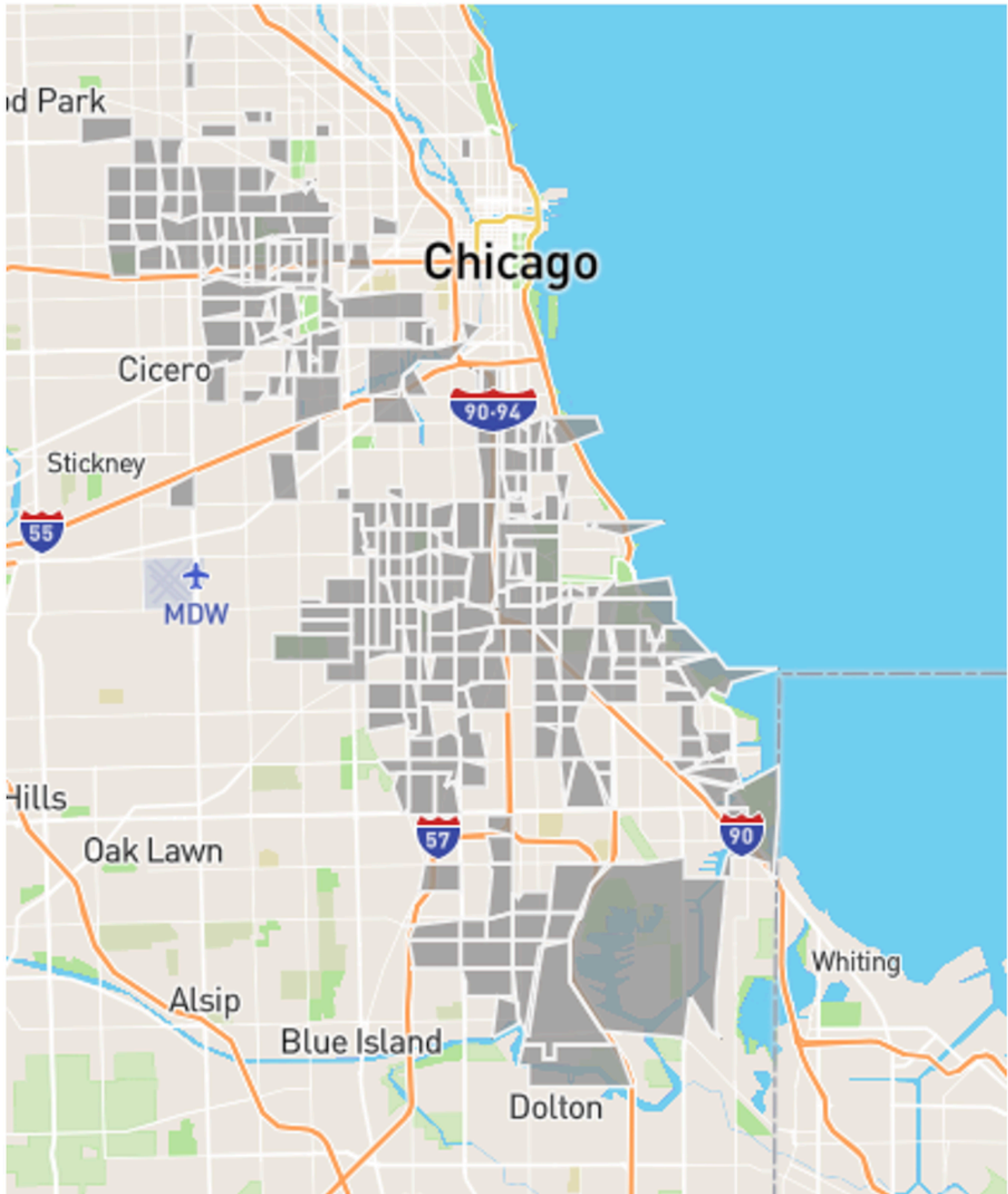
Steps Completed Before Modeling Data:

- Geographic analysis
- Prepared datetime data for analysis
- Merged the DataFrames for modeling purposes

Modeling Progression:

- Created a baseline model >> Due to the severe imbalance, a dummy model that optimizes using majority class results in 98.2% accuracy, but 0% recall and an AUC score of 50% that is no better than random guessing.
- Created a first simple model - logistic regression with three predictors >> This model does not successfully predict any positive cases, although the higher AUC suggests setting a lower probability threshold might lead to better predictions.
- Created a logistic regression model with many predictors which must be encoded >> This improved the ROC-AUC marginally, but the recall is still 0. The class imbalance is too great.
- Created a logistic regression model with SMOTE >> The final complex model is a very good model which finally captures 73% of true positives. There are many more false positives as well, which is a predicted effect of improving recall. In this case, it is an acceptable trade-off.

|	Severe Crash Not Predicted	Severe Crash Predicted
Severe Crash Did NOT Happen	113,927	46,437
Severe Crash DID Happen	800	2,145

Data Analysis Results

Our analysis shows that the following factors seem to have the largest effect on fatal or serious injury car accidents:

- The use of safety features
- The driver's condition
- Whether the crash happened during the night
- Whether the crash happened in a socioeconomically disadvantaged area (SDA)

Below, the coefficients are shown and calculated odds increase of each factor:

Feature	Odds Incr
Poor Safety	678.3%
Incapacitation	228.7%
Night-time (v. Midday)	90.0%
SDA	28.4%
Morning Rush (v. Midday)	10.2%
Summer (v. Spring)	9.7%
Evening Rush (v. Midday)	7.9%
Autumn (v. Spring)	5.8%
Weekend	5.6%
Bad Weather	4.0%
Slippery	-0.5%
Winter (v. Spring)	-2.1%
Curves / Unlevel	-8.1%
Poor Visibility	-8.4%
Malfunctioning Stoplights	-18.1%

The effect on the odds of a severe accident is plotted below:
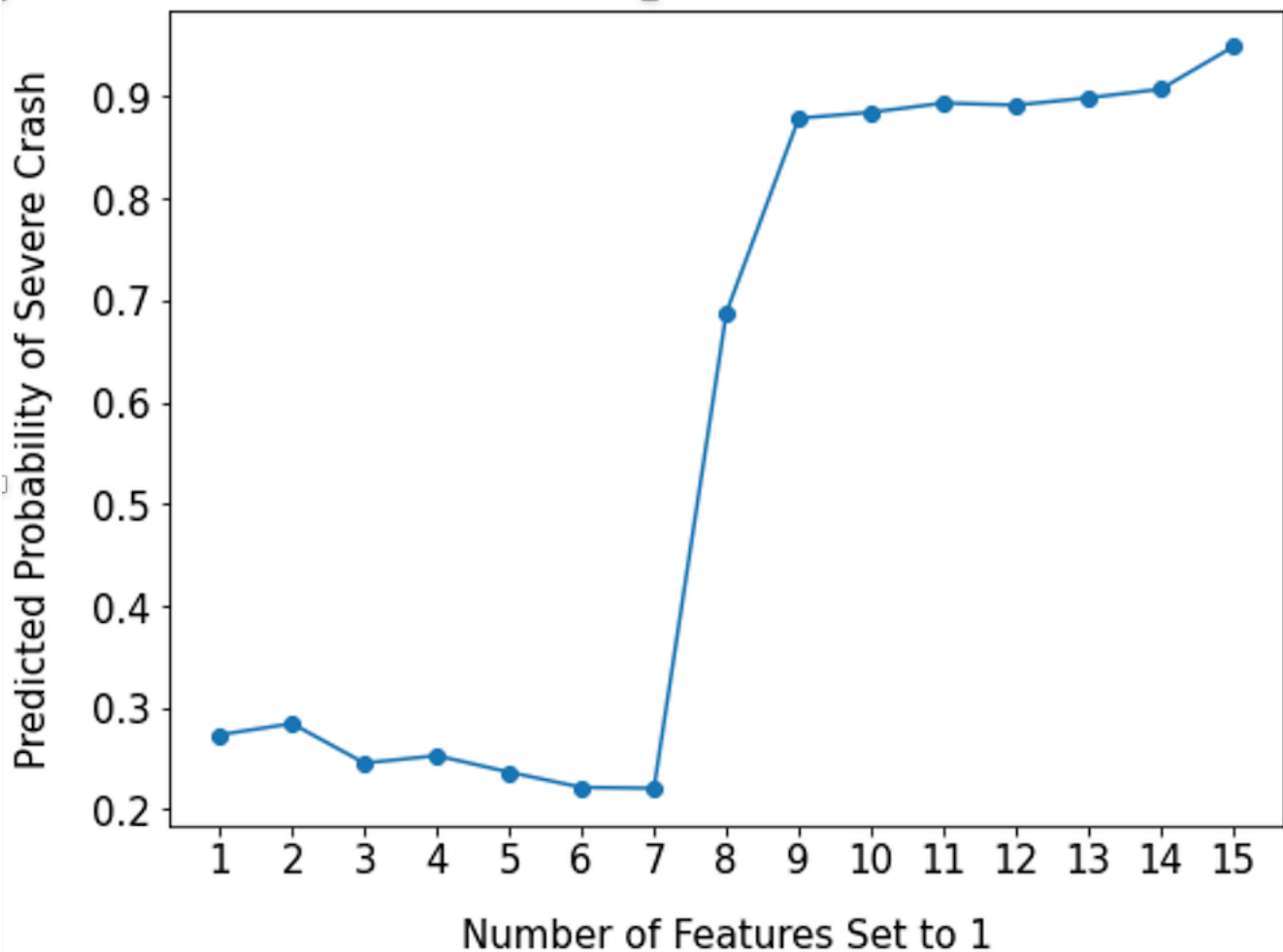
Recommendations

1. Neglecting safety measures, driving while incapacitated, night-time driving, and crashes in disadvantaged areas all significantly influence fatal or serious injury car accidents in Chicago. The data showed that incapacitated drivers, remain one of the leading factors contributing to car crash fatalities or serious injuries.

- "Every day, about 37 people in the United States die in drunk-driving crashes — that's one person every 39 minutes." - [How Alcohol Affects Driving Ability](#)

- **We propose investing in tech. With the financial resources invested, you could incentivize drivers to use devices that detect for distracted driving and other dangerous driving behaviors like speeding.**

2. We discovered a non-driving-related factor contributing to severe car accidents: being in a Socioeconomically Disadvantaged Area at the time of the accident. The diminished quality of services, including limited EMS accessibility and lower-quality hospital care in these areas, may lead to increased fatalities or serious injuries.

- **We recommend prioritizing these areas when it comes to the maintenance of traffic lights, visible crosswalks and other street design safety features.**

3. Certain measures can be implemented to decrease the likelihood of serious crashes at night. We know that a driver's vision is severely limited at that time of day more than any other time.

- **We recommend ensuring that street lamps and reflective tapes are regularly maintained to increase visibility when it is at an all time low.**

Future Investigations

- Age of the driver
- Proximity of the crash to major holidays

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors   2

rjlatail

karisteph

Languages

- **Jupyter Notebook** 100.0%