Predicting NYC Auto Collision Volume

Team Information

Team Number: 45

Jeff Hedberg

Lisa Chille

Nick Cunningham

Brittany Lange

Delband Taha

Project Background

While motor vehicle collisions have emotional, financial and health costs associated with them, they also present an opportunity to add/restore value in the form of auto repairs.

This project aims to help a fictitious local auto repair center in New York state, estimate the volume of incoming vehicles they can expect each year, assuming their market share[1] is known.

Problem Statement - Nick

Primary

- Use historic auto collision data from NYC to estimate the future collision volume.
- Then, take these collision estimates and combine them with known market share of the fictitious auto repair business to forecast their expected auto repair volume.

Secondary

- Recommend a targeted marketing campaign for market share growth, and then forecast gains
- Evaluate the effect of COVID-19 on this industry sector.

Literature review

Regev et. al. (2022)

- Evaluated the crash risk of drivers of different ages and genders adjusting for travel exposure
- Found that drivers between the ages of 21 and 29 carry the highest level of risk for being in a car crash contrary to popular belief
- Impactful for our project since it grounded us and allowed us to let the data speak

Literature review (cont.)

He et. al. (2021)

- Developed a technique using GPS data to predict where accidents happen.
- Allow policymakers to create targeted solutions e.g., install traffic calming devices in areas that are predetermined to be risky.
- We were inspired by the advanced models that they built.

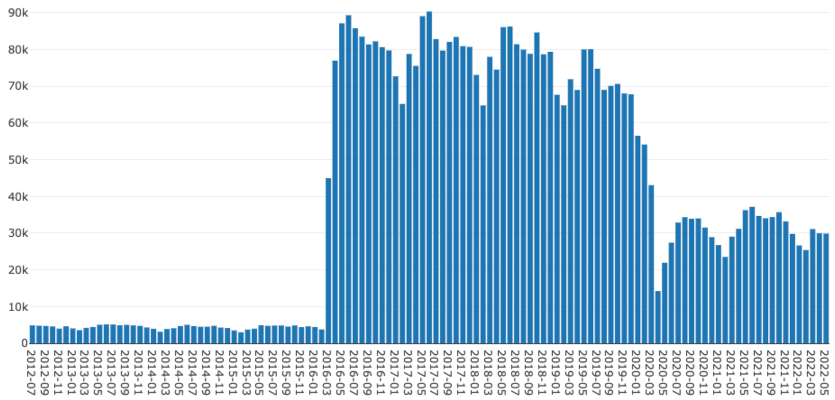
Data Sources

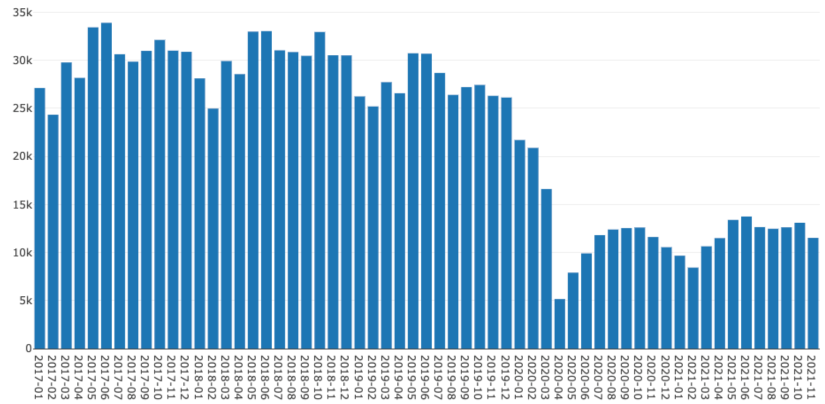
DATA.GOV

- The main data sources for this project are sourced from data.gov. These data represent New York state vehicle collisions. We will be using data from 3 sources, which share common factors: COLLISION ID, UNIQUE ID, and VEHICLE ID.
    - The Crashes dataset contains data on the actual crash events in NYC from 2013-01-01 to 2021-12-31.
    - The vehicle dataset contains data on the actual crash vehicles in NYC from 2013-01-01 to 2021-12-31.
    - The person dataset contains data on the actual crash persons in NYC from 2013-01-01 to 2021-12-31.

Known Issue

- In our research, we discovered that as a result of a traffic safety initiative to eliminate traffic fatalities, the NYPD replaced its record management system with an electronic one, (FORMS), in 2016.*** We see this change reflected in the chart to the right. The amount of data collected from March 2016 on greatly surpasses the amount previously collected.
- Resolution:
    - Because of this, and the fact that we want to help our local repair center forecast their yearly volume, we will use data collected after 2017-01-01 for full year modeling





Approach

STEP 1 — Data Wrangling

Join datasets, engineer synthetic features

STEP 2 — Feature Selection

Use identified key predictors, experiment w/ LASSO elastic net for auto feature selection

STEP 3 — Training/Validation

Cross-Validate/Train using different models, select based on R^2 or MAE

STEP 4 — Testing

Compute expected car crash count on test data using best model

FINAL — Prediction

Use model to estimate demand for local shop * market share

Hypotheses

1. Be able to estimate the monthly demand for a fictitious auto shop in New York state by extrapolating the monthly car crash volume with reasonable accuracy.

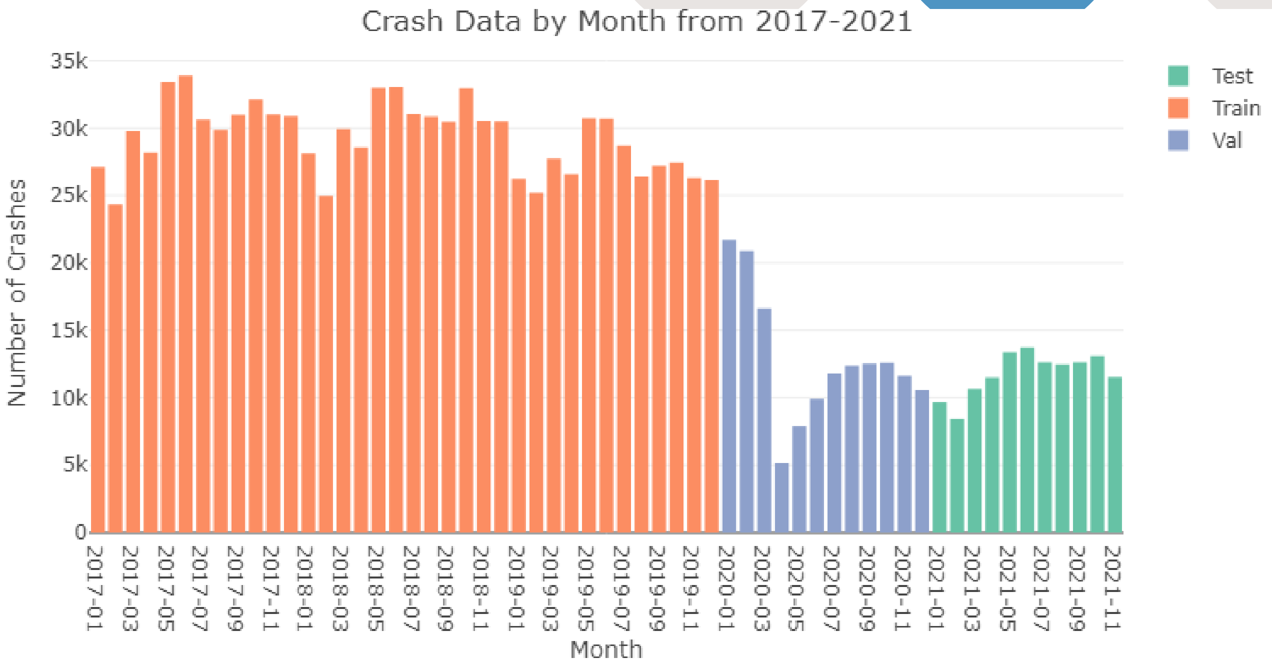
2. Identify a demographic that comprises the largest market segment for motor vehicle repairs after an accident; from our preliminary investigation, we suspect that young men comprise this demographic.

3. Be able to assert that $x expenditure on advertising targeted towards the most valuable demographic will result in $x+y revenue.

4. Be able to quantify the impact of COVID-19 on demand in this sector. Our preliminary investigation has already highlighted this.

Data Partitioning and Modeling Steps

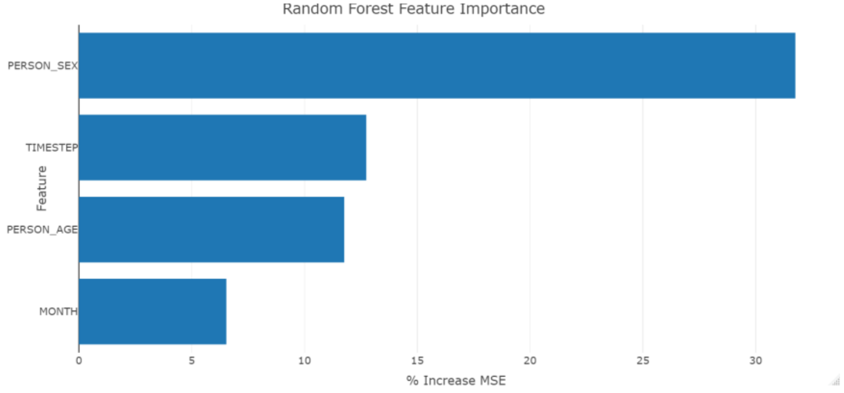
- Time-aware Modeling required chronological partitioning
- Initial Models / Hyperparameter Tuning
  - Training – 2017-01 to 2019-12
  - Validation – 2020-01 to 2020-12
- Models Rebuilt on Train + Validation
  - Training + Validation - 2017-01 to 2020-12
- Test set left aside for Final evaluation



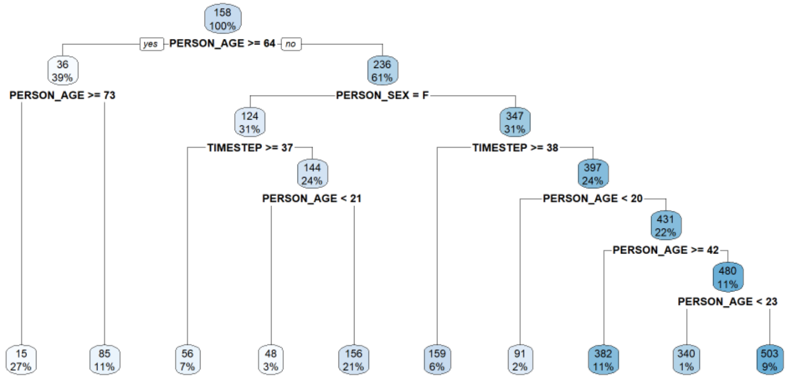
Model Evaluation Overview

- Models Explored:
  - Multiple Linear Regression
  - Random Forest
  - Classification and Regression Trees
  - Extreme Gradient Boosted Trees
  - Ensemble of models above

Random Forest model



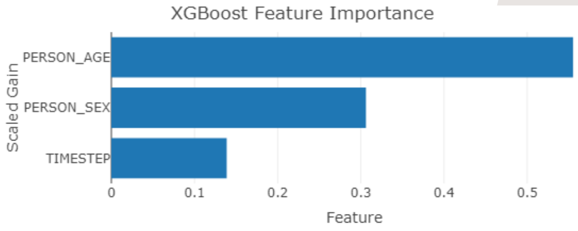
CART Model



Linear Model

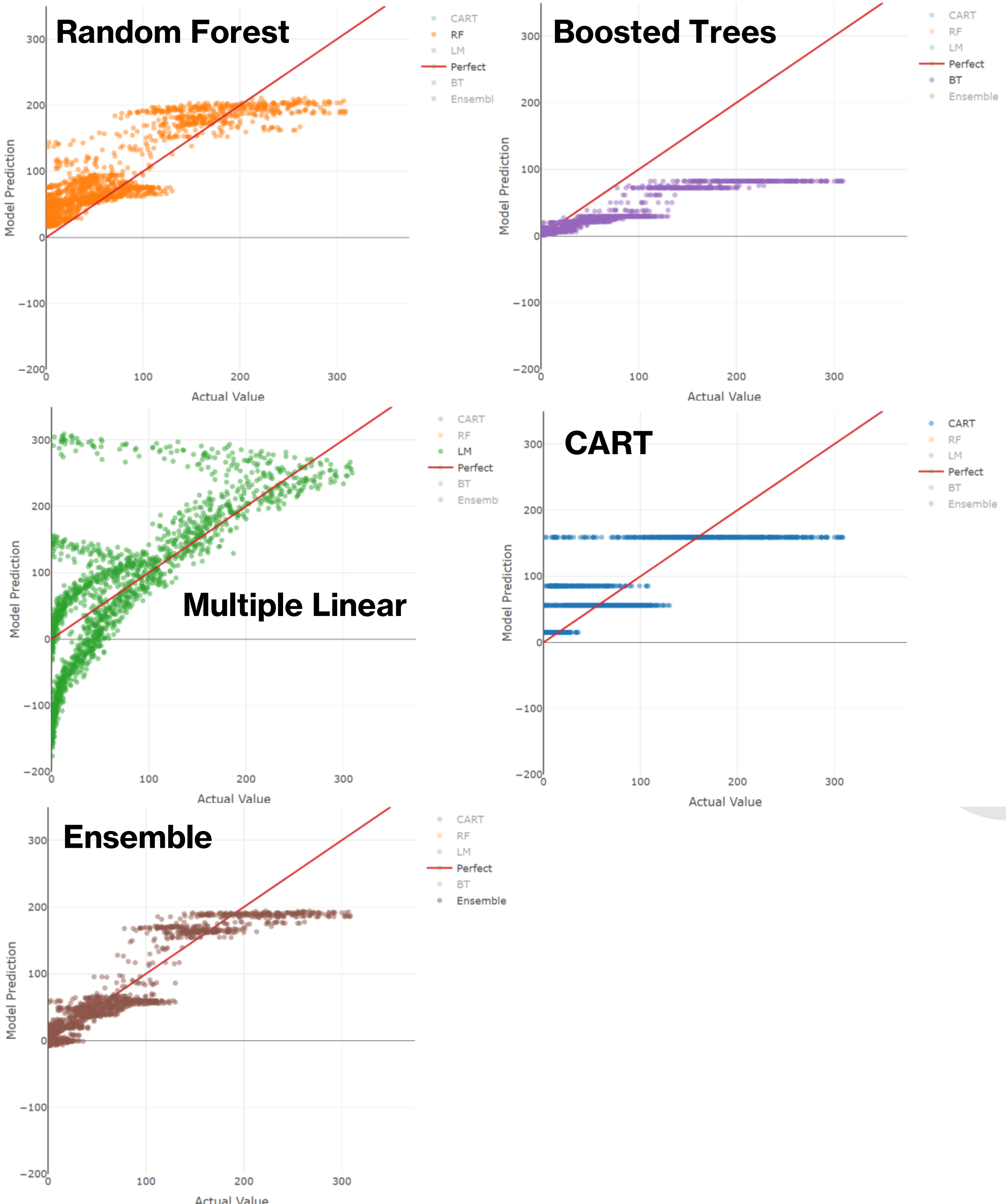
```
Call:
lm(formula = n ~ TIMESTEP + MONTH + PERSON_AGE + PERSON_SEX,
    data = train_val_df)

Residuals:
    Min      1Q  Median      3Q     Max
-441.83  -64.83    4.23   58.78  299.83

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 352.82536    5.76803  61.169  < 2e-16 ***
TIMESTEP     -2.92533    0.09545 -30.646  < 2e-16 ***
MONTH02     -10.16349    6.23071  -1.631 0.102891
MONTH03       5.85028    6.23281   0.939 0.347952
MONTH04     -17.30053    6.27381  -2.758 0.005837 **
MONTH05      13.68098    6.25907   2.186 0.028862 *
MONTH06      19.14640    6.25344   3.062 0.002208 **
MONTH07      13.20477    6.25124   2.112 0.034689 *
MONTH08      12.80773    6.26345   2.045 0.040905 *
MONTH09      18.11904    6.27356   2.888 0.003886 **
MONTH10      27.72251    6.28432   4.411 1.04e-05 ***
MONTH11      22.61087    6.30316   3.587 0.000336 ***
MONTH12      23.03289    6.32485   3.642 0.000273 ***
PERSON_AGE   -3.92659    0.05504 -71.344  < 2e-16 ***
PERSON_SEXM 150.90275    2.55012  59.175  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 111.3 on 7619 degrees of freedom
Multiple R-squared:  0.5457,    Adjusted R-squared:  0.5449
F-statistic: 653.8 on 14 and 7619 DF,  p-value: < 2.2e-16
```

XGB Model



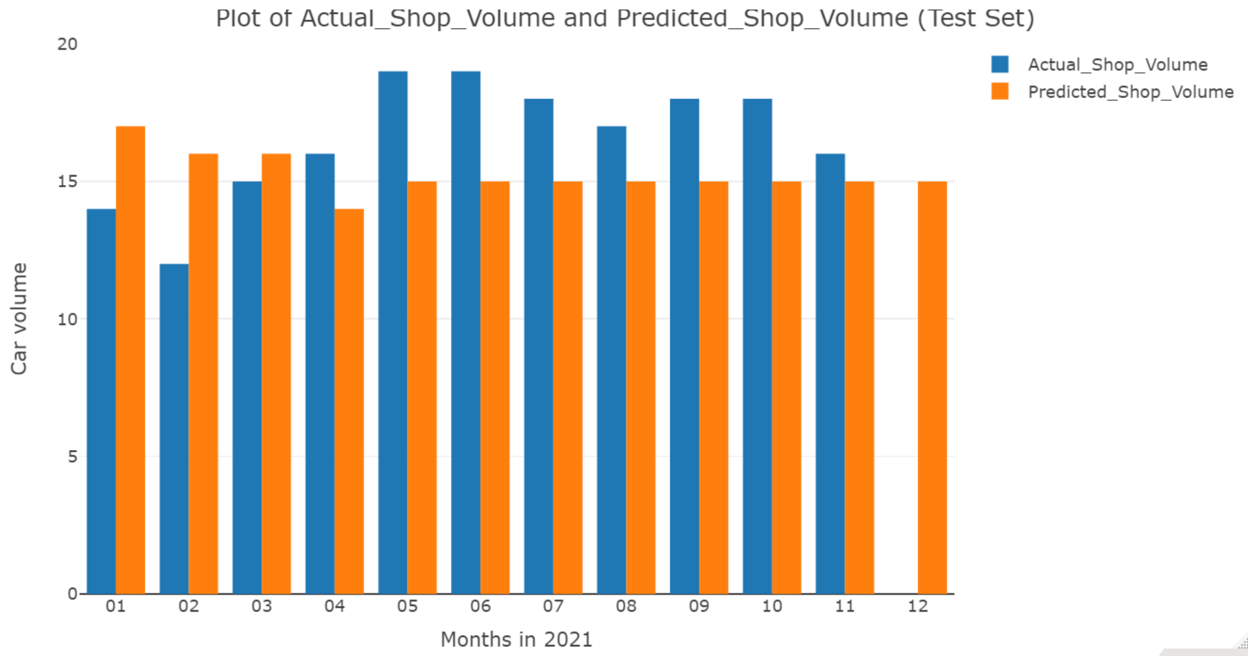
Modeling Results

Model_Type	R_sq_train	R_sq_val	R_sq_train_val	R_sq_test
Linear	0.58	-1.15	0.55	0.08
Random Forest	0.87	-0.49	0.85	0.73
CART	0.95	-1.75	0.90	0.65
Boosted Trees	0.37	0.77	0.41	0.23
Ensemble	NA	0.91	0.96	0.86



Ensemble Model Output

- Model performs very well
- We hypothesize that the model would be even better with out the impact of COVID-19.
- Next, we'll discuss secondary research objectives for our project.



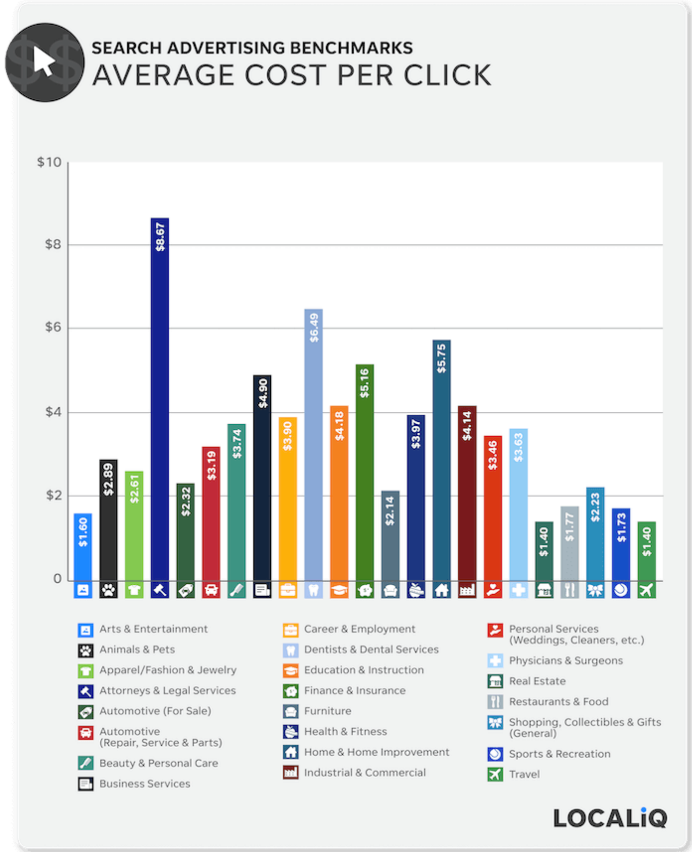
Secondary Problem Statement: Marketing Campaign

- Recommend a targeted marketing campaign for market share growth, and then forecast gains

2021 Paid Search Industry Benchmarks

Automotive — Repair, Service & Parts

CTR	5.39%
CPC	$3.19
Conversion Rate	15.23%
Cost-per-Lead	$17.81



Marketing Campaign – Financials

- In-order-to build out a costing model of shop performance, we sourced average financial metrics from an industry publication. We adjusted the ARO to $2400 to account for collisions being a higher cost service.

Average Repair Shop Financial Metrics

Average Repair Order (Units)	$2,400
Gross Profit	40%
Net Profit Margin	10%

The Typical Shop

Just over 200 industry professionals completed the Shop Performance Survey, and, while they were evenly dispersed across all U.S. markets, the majority of respondents followed a distinct demographic pattern that also closely aligns with our overall readership.

The Average Shop

Shop Type: Independent repair business (87%)

Work Type: General repair (68%)

Shop Size: 2,000–4,999 square feet (41%)

Number of Lifts: 3–4 (32%)

Number of Bays: 3–4 (28%)

Annual Revenue: $1M–$2.49M (30%)

Average Monthly Car Count: 100–149 (19%)

Average Repair Order: $200–$399 (45%)

Gross Profit Margin: 40–49% (28%)

Net Profit Margin: 10–14% (22%)

Marketing Campaign - Performance

- Using the Industry Search Benchmarks and Financial metrics we were able to cost out performance of three different Search Ad Spends.
- If we have a Market Share Goal of reaching 25 basis points, we can see that a $300 monthly ad campaign gets closest to achieving this.

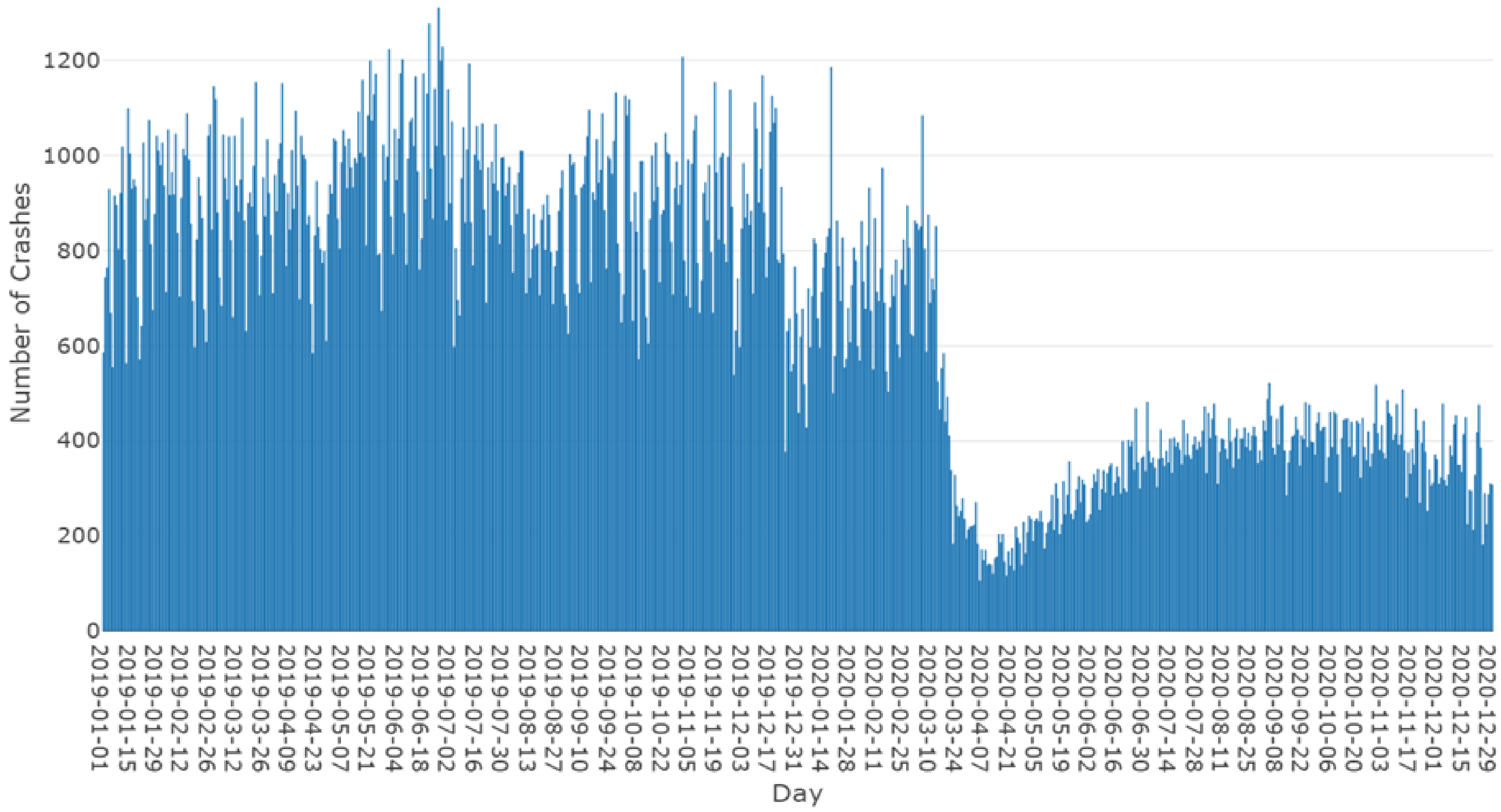
	Search Ad Spend		
Campaign KPIs	$200	$300	$400
Clicks	63	94	125
Conversion	10	14	19
Revenue	$22,916.61	$34,374.92	$45,833.23
ROAS	114.58	114.58	114.58
Market Share Lift	0.0006	0.0009	0.0012
Gross Profit	$9,166.65	$13,749.97	$18,333.29
Net Profit	$2,291.66	$3,437.49	$4,583.32
ROI	1146%	1146%	1146%

Assumes cost/unit static w/ increase (labor + parts)

Secondary Problem Statement: Effect of COVID-19

- First reported case of COVID-19 in New York state on March 1, 2020, but as many as 10,700 New Yorkers had already contracted the virus
- March 9: 16 cases in NYC alone
- March 16: public schools close, beginning COVID-19 lockdown
- Stay-at-home orders in place until June 8, when phase 1 of reopening began under safety protocols
- Estimated 44% of all metro NY residents infected by end of 2020
- Total of 25,000 confirmed deaths of NY citizens with 5,000 probable

NYC Vehicle Collision Data by Day



Regression Analysis

$$Crashes = \beta_0 + \beta_1 * Covid$$

- Using COVID as an indicator variable and the number of daily crashes in NYC as the response variable, we built a linear regression model to analyze the effect of COVID-19 on vehicle collisions in NYC.
- Our findings suggest that before Covid, there was an average of 602 crashes a day in NYC. This number dropped to 233 crashes a day after Covid.
- NYC safer-at-home policies led to a 61.3% reduction in vehicular collisions

```
Call:
lm(formula = n ~ covid, data = covid_df)

Residuals:
   Min      1Q Median      3Q     Max
-600.8 -230.8  102.7  245.5  708.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   601.80      13.77   43.71   <2e-16 ***
covid        -369.04      21.73  -16.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 347.5 on 1062 degrees of freedom
Multiple R-squared:  0.2135,    Adjusted R-squared:  0.2128
F-statistic: 288.3 on 1 and 1062 DF,  p-value: < 2.2e-16
```

Decrease in Costs Due to Accidents

13.6 million car accidents annually in United States	The average cost of vehicle repair per collision: $21,036
22% of collision claims involve vehicles that are totaled	Accidents not totaled will be potential business for our auto body repair center
470 average accidents/day before COVID-19	182 average accidents/day after COVID-19

Effect on Fictitious Auto Body Repair Shop

$6,054,581 in lost revenue for auto body shops in NYC post COVID

Assuming .14% market share for our auto body shop, approximate lost revenue totals $8,476 a day

Total of $712,019 lost revenue during the COVID-19 lockdown in NYC from March - June

Literature review (cont.)

Li and Zhao (2022)

- Quantity of vehicular collisions has plummeted, but cyclists’ fatalities have tripled since the start of the pandemic.
- Cycling kilometers increased by 150% in Philadephia.
- Collisions tend to happen in temporal and spatial hot zones.

Future work

- Understand differences between traffic collisions that involve cyclists and those that do not.
- Determine features of vehicular and cycling collision hot spots.
- Use the sophisticated formulae for traffic exposure from Regev et. al. (133) on new data sets to see if it applies in other places.
- Perform more data exploration through advanced visualizations to pull out and analyze any other trends in the data.

Sources Cited

Barnes, Stephen, et al. "COVID-19 Lockdown and traffic accidents: Lessons from the pandemic." *Contemporary Economic Policy* 40.2 (2022): 349-368.

Bean, Travis. *The 2017 Shop Performance Survey*. 01 November 2017. July 2022. <https://www.ratchetandwrench.com/articles/5258-the-2017-shop-performance-survey>.

CDC COVID-19 Response Team. "Timing of State and Territorial COVID-19 Stay-at-Home Orders." *Morbidity and Mortality Weekly Report* 69.35 (2020): 1198-1203.

City of New York. *NYC Health*. 1 April 2020. 14 July 2022. <https://www1.nyc.gov/site/doh/covid/covid-19-data.page>.

He, Songtao, et al. "Inferring high-resolution traffic accident risk maps based on satellite imagery and GPS trajectories." *IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal: Institute of Electrical and Electronic Engineers, 2021. 11957-11965.

Lam, Lawrence. "Distractions and the risk of car crash injury: The effect of drivers' age." *Journal of Safety Research* 33.3 (2002): 411-419.

Lewis, Caroline. *Officially, NYC Had One COVID-19 Case. New Research Suggests It Was Actually 10,000*. 24 April 2020. 14 July 2022. <https://gothamist.com/news/officially-nyc-had-one-covid-19-case-new-research-suggests-it-was-actually-10000>.

Li, Jintai and Zhan Zhao. "Impact of COVID-19 travel-restriction policies on road traffic accident patterns with emphasis on cyclists: A case study of New York City." *Accident; analysis and prevention* 167 (2022).

McCormick, Kirsten. *2021 Paid Search Advertising Benchmarks for Every Industry*. 20 May 2022. July 2022. <https://www.wordstream.com/blog/ws/2021/10/13/search-advertising-benchmarks>.

Sources Cited (cont.)

New York State Department of Motor Vehicles. *Find a DMV Regulated Business*. n.d. June 2022. <https://process.dmv.ny.gov/FacilityLookup/?_ga=2.33023101.270841510.1656975631-932475252.1656975630&TSPD_101_R0=084c043756ab200049c13ac02223efa9441583bad77008d3b71
7 90bcb79ba757be5b2aa10507b29f084bbc5efb143000172ce93dfec8420e0dc96ad9ed441e7f4a2448770215c0d>.

Police Department (NYPD). *Motor Vehicle Collisions - Crashes*. 19 July 2022. 30 May 2022. <https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95>.

—. *Motor Vehicle Collisions - Person*. 29 June 2022. 30 May 2022. <https://data.cityofnewyork.us/Public-Safety/Motor- Vehicle-Collisions-Person/f55k-p6yu>.

—. *Motor Vehicle Collisions - Vehicles*. 8 December 2021. 30 May 2022. <https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles/bm4k-52h4>.

Regev, Shirley, Jonathan Rolison and Salissou Moutari. "Crash risk by driver age, gender, and time of day using a new exposure methodology." *Journal of Safety Research* 66 (2018): 131-140.

Thompson, Karl. *What is Normal?* 9 September 2018. July 2022. <https://revisesociology.com/2018/09/03/what-is-normal/>.

Vallet, Mark. *Total loss thresholds by state*. 16 September 2021. 17 July 2022. <https://www.carinsurance.com/Articles/total-loss-thresholds.aspx>.

Volovich, Kristina. *What's a Good Clickthrough Rate? New Benchmark Data for Google AdWords*. n.d. July 2022. <https://blog.hubspot.com/agency/google-adwords-benchmark-data>.

Williams, Allan. "Teenage drivers: patterns of risk." *Journal of Safety Research* 34.1 (2003): 5-15.

Thank you for watching our presentation!

Team Information

Team Number: 45

Jeff Hedberg

Lisa Chille

Nick Cunningham

Brittany Lange

Delband Taha