Training AI to Detect Stolen Car Parts Online:

Research and Risk Score Formula

Harmond Drenth, Kurt Wokoek, Warren Burrus, Courtney Hodge, Denny Lee, and Tacoma Velez

Principal Investigator: Dr. Pablo Rivas

Baylor University

Research Findings

Carver 2014

The 2014 research paper by Christopher Carver outlines different criteria for identifying suspicious online posts. These criteria are categorized as primary or secondary red flags within posts, and they can determine how.

Red Flags

Primary red flags are strongly indicative of how illicit a post may be, and include:

- The date and time of the posting will be relatively close to the time the item was stolen.
- The seller is in close proximity to where the item was stolen.
- The item is being sold for less than the average market price.

Secondary red flags are less indicative of an illicit post but are still significant, and include:

- The post is using a stock photo or photo from another post for the item they are selling.
    - Either not displaying a photo of the item or displaying a stock image instead.
- The seller either negotiates or displays the desire to sell the item away from their residence.
- The post contains a poor description of the item and the seller does not appear to have much knowledge of the item.
- The post indicates that the seller is overeager to sell the item.
- The post contains a telephone or contact information for the seller that is spelt out or obfuscated rather than in plain digits.

Measuring Red Flags

Primary red flags can be systematically measured as follows:

- **Date and Time** - measure the difference in time between the time of theft indicated by a reported case and the time the post was created.
- **Proximity** - measure the difference in location of theft and seller, however only using city granularity (thieves will sell item in surrounding area).
- **Price** - measures the difference between the posts price and the average market price of the item (suspicious posts fall below average).

    - The suspiciousness of the posts price can be defined by $1 - \frac{1}{k|\Delta p|}$ where $\Delta p = \frac{|P_{Market} - P_{Post}|}{P_{Market}}$ is the variance in price from the market average, allowing all posts to have suspiciousness but only relative to their difference from the market average.
    - k represents a domain specific designed to conform the suspiciousness of the posts to that standard deviation from the market average.
    - Problem with $1 posts; but could automatically be suspicious, when combined with other factors

Secondary red flags can be systematically measured as follows:

- **Stock Photos** - take a hash of seller's image (if provided) and compare it to hashes of popular Google image results for that classification and other posts images.
- **Poor Description** - compare the seller's description to commonly marketed descriptions, to find how much of the description overlaps with other posts (unique descriptions are less suspicious than copy/paste).
- **Eager to Sell** - small common phrases such as "want to sell fast", "want to sell as soon as possible", "need cash quick", etc.
    - Using short 3-5 n-gram common strings, we can attempt to identify this attribute; furthermore, these strings could be compiled automatically from suspicious posts.
    - To find keyword trends, search for n-gram phrases that were common amongst all suspicious posts.
        - But removing the n-gram requirement would allow for catchy keywords or other aspects to be discovered.
- **Contact Info** - inconsistencies such as different seller names or telephone numbers for posts within the same timeframe are highly suspicious.
    - Ex: obfuscate their contact information by spelling it out or adding format characters to prevent search engines from effectively clustering their posts.
    - This can be achieved by indexing the seller's full name (if provided) to their contact information along with a date range of use, if there are many overlapping ranges with different contact information then it is highly suspicious.

Theft-Reporting Database

The paper also discussed using a Theft-Reporting Database, which allows the public or police authorities to submit the stolen reports into a linguistically fixed database. Users would be asked to select or fill in the information regarding their stolen property, such as stolen item, make, model, color, etc.

As an alternative to an ongoing database, we also propose scraping police reports or news articles, or calling APIs to police and news sites on thefts. Suspicious posts from online platforms could be cross-referenced with thefts reported by police and news, to determine if any posts correspond to a crime.

ORC Report

The ORC (Organized Retail Crime) Report outlined additional red flags related to illicit online posts, primarily on how physical proximity to high-crime cities suggests suspicious posts. The report also elaborated on some of the same red flags from Carver 2014, namely related to the post's price, photos, and description.

Red Flags

The 10 cities with the highest ORC volume:

- 1 New York
- 2 Los Angeles
- 3 Philadelphia
- 4 Chicago
- 5 DC/Baltimore
- 6 San Francisco
- 7 Houston
- 8 Miami
- 9 Dallas
- 10 Orlando/Tampa

The report also further explained other red flags in online marketplaces:

- **Price**
  - Item price is significantly less than price of other sellers on the marketplace or below manufacturer's cost.
- **Photos**
  - Merchandise still in shipping plastic (cargo theft).
  - Large variety of sizes available.
  - A variety of merchandise, new with tags.

    - Different sellers using the same photos, or posting photos of merchandise with the same background.
    - Merchandise photos with sensor tags or other electronic article surveillance (EAS) devices still attached.
    - Defaced product labels.
    - Using stock retail photos.
    - Photos of merchandise taken inside a vehicle.
- **Description**
    - Specific language using words such as: “like new”, “new in box” or “NIB”, “new with tags” or “NWT”, “unopened”, “taking orders [for product]”, “DM for orders or size”, “factory sealed”.

Risk Score Formula

Overview

Using our research from Carver 2014 and the ORC Report, we derived a risk score formula to systemically rate the potential riskiness of a given post on a scale from 0.0 (not risky) to 1.0 (very risky).

This risk score formula is composed of seven different parameters, five of which are based on the red flags outlined in our research: a post’s (1) price; (2) photos; (3) description; (4) contact information; and (5) crime city proximity. Additionally, our formula includes two other parameters recommended by our client: (6) duplicate posts and (7) similar timestamps. These two parameters are intrinsically linked and apply only to similar or identical posts spread across platforms.

Weights

The formula calculates the total risk score by assigning weights to each of these parameters, calculating an individual score for each parameter, and summing all parameter scores together.

Category	**Weight**
Price	0.25
Photos	0.10
Description	0.15
Contact Info	0.20
Duplicate Posts	0.10
Similar Timestamps	0.10
Crime Cities	0.10

Parameters

Each parameter and how to compute it are described below:

Price

$$0.25\left(\frac{|P_{Market} - P_{Post}|}{P_{Market}}\right)$$

- **P$_{\textbf{Market}}$** = average market price of car part
- **P$_{\textbf{Post}}$** = selling price of car part in post

This parameter measures the percentage below the standard market value of the post's price, with the 40-75% range being risky.

This parameter has not yet been coded but could be automatically calculated in future development by referencing eBay, Craigslist, OfferUp, or similar online marketplaces for their average market price of a given car part.

Currently, P_{Market} is calculated as the maximum P_{Post} of all user prices for that car part in the dataset, based on findings that illicit sales will sell parts for less than the market price, making the highest price the least suspicious and closest to the actual market price.

Photos

$$0.10\left(1 - \frac{1}{R+1}\right)$$

- **R** = number of **R**ed flags in photos (tags, opened packages, etc.)

This parameter measures the number of red-flag features in photos—any suspicious imagery or metadata in the photos as outlined in our research, like unopened packages or EAS devices.

This parameter has not yet been coded due to the difficulty of automatically identifying suspicious clues in photos. This parameter could be implemented with image-reading AI trained on illicit car part images, or by analyzing metadata of photos.

Description

$$0.15\left(1 - \frac{1}{K+1}\right)$$

- **K** = number of suspicious **K**eywords, phrases and terms in description

This parameter measures the number of red-flag phrases and terms in post descriptions, as outlined in our research. Examples include phrases with an eagerness to sell, like “want to sell fast” or “need cash quick”; and terms like “new in box”, “unopened”, or “DM for orders or size”.

This parameter is implemented programmatically by using BLEU score and semantic score analysis in bleuAndSemantic.py. Each sentence of a post is given a BLEU score and semantic score based on whether it matches a given reference bank of red-flag phrases.

Any sentence whose scores meet or exceed the threshold values are counted as red flags themselves, providing a total count of red flags for the post. Early testing suggests a BLEU threshold of around .68 or above, which corresponds to industry consensus, and a spaCy semantic threshold of around .75. Future work could be done in refining the phrase list, and even potentially detecting different categories of red flags and weighing them differently.

Contact Info

$$0.20\left(\frac{N+P}{2}\right)$$

- **N** = binary 0 or 1 if **N**ame is standard first and last name or not, respectively
- **P** = binary 0 or 1 if **P**hone number is standard number format or not, respectively

This parameter is a binary measure of whether a post contains an unusual username or phone number, with “yes” being 1 and “no” being 0 for each. An unusual name would be different from a standard first and last name, like an alias. An unusual phone number is partially or entirely spelled out, or different from the standard number format, to avoid regular expression matches.

This parameter is implemented programmatically in ResearchQuestions.ipynb as an answer to a research question on phone numbers. It uses REGEX pattern analysis to identify suspiciously spelled out phone numbers indicative of illicit sales.

Duplicate Posts

$$0.10\left(1-\frac{1}{D+1}\right)$$

- **D** = number of **D**uplicate posts with same content

This parameter measures the number of duplicates of posts within and across platforms like Craigslist or OfferUp. Posts are considered duplicates if they contain similar or the same content.

This parameter has been programmed to detect any duplicate posts. Duplicate or similar posts are found by comparing the text of the post bodies. First, strip all non-alphanumeric characters from the post body, then use the stripped body to generate a hash of the post, stored in a hash set with a reference to the row. If we've seen the hash before, mark both the original row and row we've just read as duplicates. Continue with all posts in the dataset (1.5 million in just one of these datasets). When we've read all the posts, write all the duplicate posts and their groupings to a duplicates file processed_data_offerup_v2_duplicatesonly.csv. Any non-duplicate posts get written to a deduplicated file processed_data_offerup_v2_deduplicated.csv.

Timestamps

$$0.10\left(\frac{1}{\Delta T + 1}\right)$$

- **ΔT** = difference in days between **T**imestamps of two posts, or between a post and average timestamp of all its duplicates

This parameter measures the difference in days between timestamps of the duplicate or similar posts identified in the above Duplicates parameter. As such, the Timestamps parameter applies only to duplicate posts and is derivative of the Duplicates parameter; if Duplicates is 0, then Timestamps will naturally be 0, as well.

The closer the timestamp of each similar or duplicate post is to the average of all similar or duplicate posts, the riskier and more suspicious they become.

This parameter has been implemented within RiskScore.ipynb for testing the risk score formula. A function and code block read the timestamps of all duplicated posts and compare each timestamp to their collective average time.

Crime Cities

$$0.10\left(\frac{100 - r}{100}\right)$$

- **r** = mile **r**adius from high-crime city (0 <= **r** <= 100)

This parameter marks posts from or near high-crime cities (NYC, LA, etc.) as risky, whereby a closer proximity to such cities garners a higher score. Per our client’s recommendation, we assume crime networks to operate within a 100-mile radius of high-crime cities, so posts outside this range are scored a zero.

This parameter has been partially implemented with geo-plots. Area codes are extracted from the phone numbers of posts and cross-referenced with an area code website to determine proximity to high crime cities plotted in geo-plot maps. Future development could query Google Maps API to determine a post location’s exact mileage from the nearest high crime city.

Complete Formula

The combined formula sums the individual scores of each parameter, producing a post's total risk score on a scale from 0.0 (low) to 1.0 (high).

$$0.25\left(\frac{|P_{Market} - P_{Post}|}{P_{Market}}\right) + 0.10\left(1 - \frac{1}{R+1}\right) + 0.15\left(1 - \frac{1}{K+1}\right) + 0.20\left(\frac{N+P}{2}\right)$$

- **P$_{\text{Market}}$** = average market price of car part
- **P$_{\text{Post}}$** = selling price of car part in post
- **R** = number of **R**ed flags in photos (tags, opened packages, etc.)
- **K** = number of suspicious **K**eywords, phrases and terms in description
- **N** = binary 0 or 1 if **N**ame is standard first and last name or not, respectively
- **P** = binary 0 or 1 if **P**hone number is standard number format or not, respectively

$$+ 0.10\left(1 - \frac{1}{D+1}\right) + 0.10\left(\frac{1}{\Delta T+1}\right) + 0.10\left(\frac{100-r}{100}\right)$$

- **D** = number of **D**uplicate posts with same content
- **ΔT** = difference in days between **T**imestamps of two posts, or between a post and average timestamp of all its duplicates
- **r** = mile **r**adius from high-crime city (0 <= **r** <= 100)

Rationale for Formula

This risk score formula was derived from the research we gathered from Carver 2014 and the ORC Report at the initial stage of our project. Below is further rationale for the formula in the context of these research articles.

Weight Justifications

Weights were assigned to each parameter based loosely on their precedence stressed by the research articles.

- **Price: 25%**
  - Comparing the asking price to market price was consistently a strong indicator of suspicion in both (Carver 2014) and (ORC Report), since criminals often sell for less than market price to sell quickly. Thus, price is weighed the most.
- **Photos: 10%**
  - Photos are common elements of online posts, but vary greatly in what they show, allowing suspicious photos to be distinguished from legitimate

    - photos. (ORC Report) outlines many different red flags that can be captured in photos. However, implementing photo detection in code is a challenge, so photos are currently weighed less.
- **Description: 15%**
    - Descriptions can contain many suspicious keywords hinting at wanting to sell fast (Carver 2014) or selling stolen products (ORC Report). However, people have distinct ways of typing that may unintentionally sound suspicious. Thus, descriptions are weighed intermediately.
- **Contact Info: 20%**
    - Every post inherently has a seller to be contacted, so contact info is consistent data to look for in a post. Red flags on contact info are easy to identify and strong indicators of suspicious posts, like spelling out phone numbers (Carver 2014). Thus, contact info is weighed heavily.
- **Duplicates, Timestamps, Crime Cities: 10% each**
    - These three parameters are relatively new parameters added upon request of the client. The timestamps parameter is contingent upon the presence of the duplicates parameter, which itself is rare among posts and not discussed as heavily within research. Crime cities are generally the most populated cities, which may have illicit sales but also many more benign sales in general. So, these three parameters are weighed less.

Suspicious Price Ranges

Based on price ranges of illicit sales studied in Carver 2014, the range of suspicious prices should be **40-75%** of average market price.

For one, prices in the $100 or $150 ranges were considered suspicious for a Blackberry averaging $210 (Carver 50). The $70-140 price range was considered suspiciously low-priced, whereas the $175-240 price range was assumed legitimate (50), meaning **47.6-71.4%** of market price is suspicious. The $1-70 price range was considered garbage data to be ignored, not suspicious (50). The lower 20% of all prices for the Blackberry were ultimately excluded from the analysis (51). Thus, everything under **40%** of the average market price should be ignored.

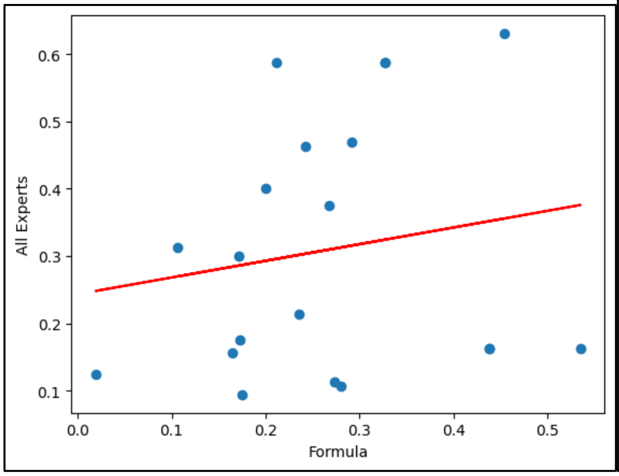
Furthermore, prices in the $125 and $150 ranges were considered suspicious for an iPhone 4 averaging $200 (53), meaning **62.5-75%** of market price is suspicious. Also, asking prices for Toronto metro pass were rarely less than 85% of market price, so not enough of a differentiation to be suspicious (80-81).

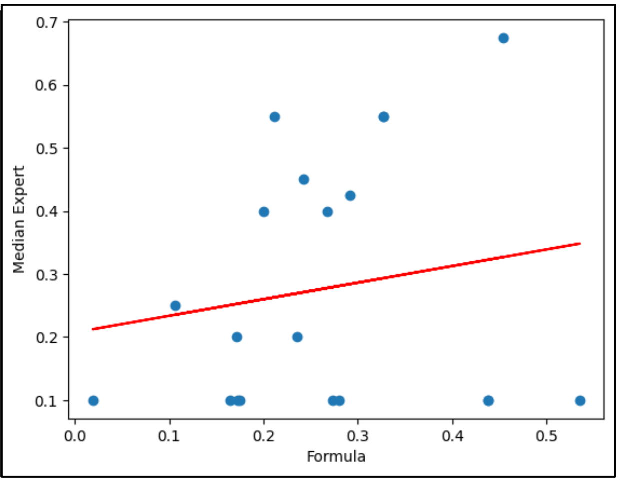
Testing Formula

Expert Feedback

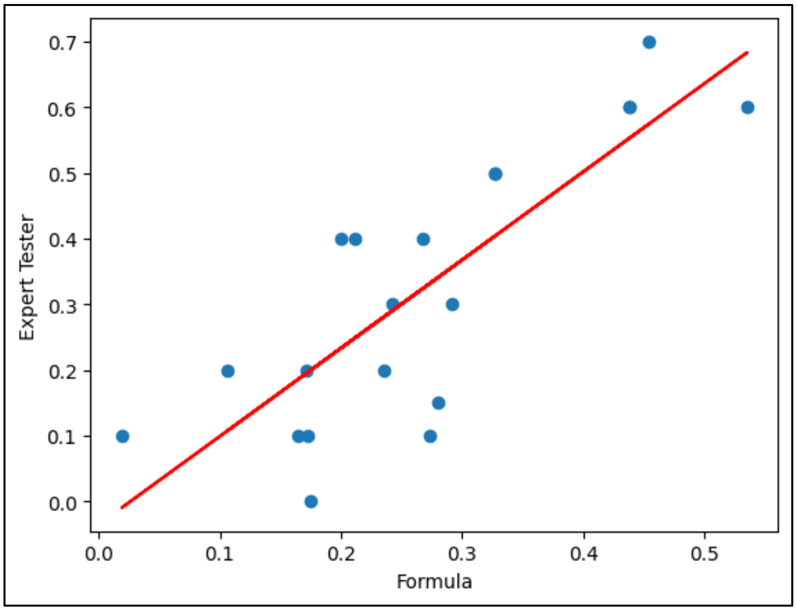
The formula's accuracy was tested by comparing its risk scores to expert scores over a set of posts. 8 experts were sampled: the principal investigator, a research assistant, and us six researchers. 17 sample posts from the dataset were selected for the 8 experts to score on the risk scale from 0.0 (not risky) to 1.0 (very risky). Results were gathered in Sample Test Scores.xlsx.

The average and median of these scores for each post were correlated with the formula's corresponding risk scores to produce a linear regression. A correlation of about +0.17 was shown to exist. This was done in RiskScore.ipynb.





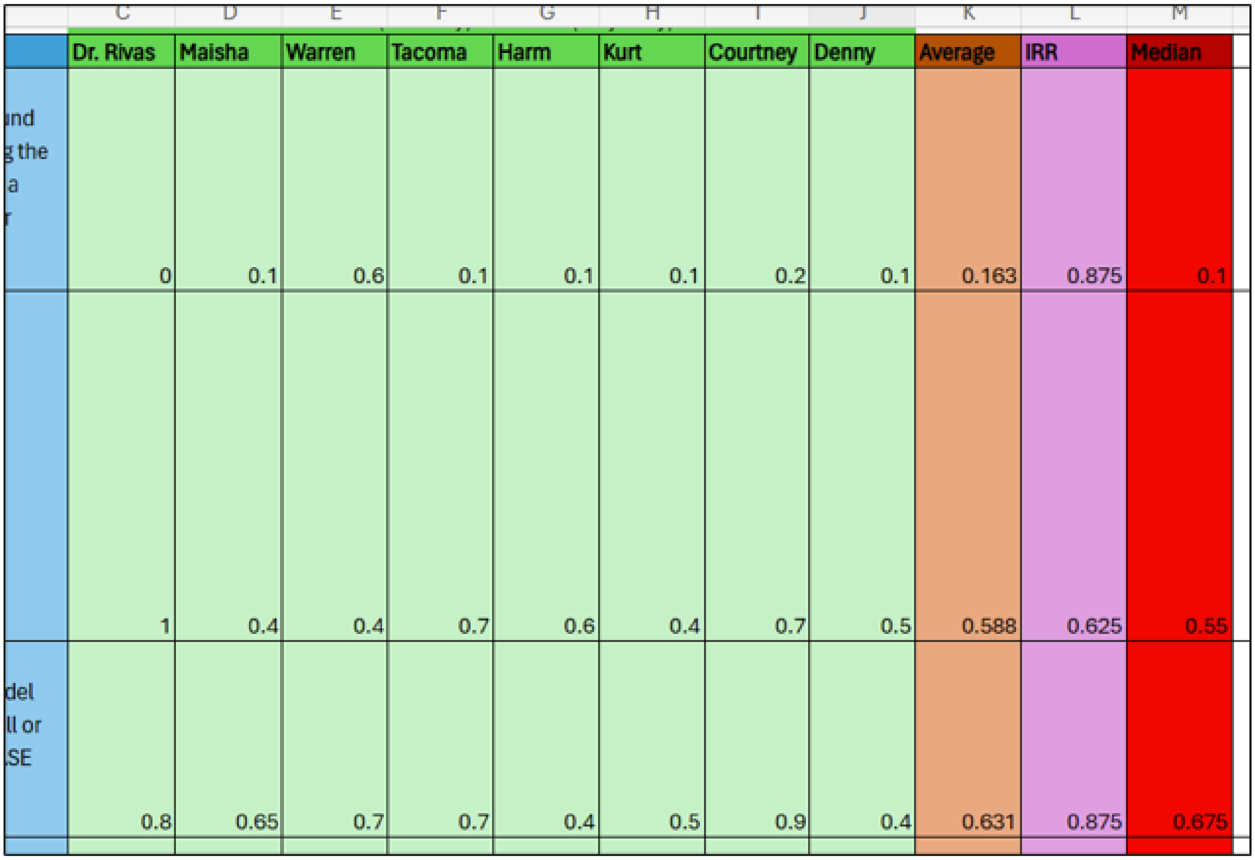
However, to raise the correlation, the formula's scores were also correlated with only the expert reviewer who primarily designed the formula, Warren. This correlation proved to be much higher, at +0.81, likely because the other experts did not evaluate or judge each post with as nuanced or fine-grained parameters as the formula does.



Inter-Rater Reliability

Inter-rater reliability was consistently strong across all 17 sample posts reviewed by 8 experts, showing general agreement on post suspiciousness with a few outliers. An IRR was computed for each post individually, measuring all reviewers' scores of that post. Every post had an IRR above 0.7 except for two, which were each 0.625; an IRR above 0.7 is a generally accepted industry standard. Since each reviewer could have slightly different scores on a ten-point scale from 0.0-1.0, the IRR was slightly modified to count all scores within a 0.2 range as in agreement. This way, all scores that fell with 0.1 of either side of a median score would agree with that score.

The IRR scores—as well as all sample posts and expert scores used for testing—are stored in Sample Test Scores.xlsx, but a subset of IRR scores are shown below.



	C	D	E	F	G	H	I	J	K	L	M
	Dr. Rivas	Maisha	Warren	Tacoma	Harm	Kurt	Courtney	Denny	Average	IRR	Median
und g the a r	0	0.1	0.6	0.1	0.1	0.1	0.2	0.1	0.163	0.875	0.1
	1	0.4	0.4	0.7	0.6	0.4	0.7	0.5	0.588	0.625	0.55
del ll or SE	0.8	0.65	0.7	0.7	0.4	0.5	0.9	0.4	0.631	0.875	0.675