>>> Keystroke Biometrics
>>> for Bot Detection

Bradley Reeves
Central Washington University
June 2, 2021

>>> Outline

1. Introduction

2. Data Collection

3. Feature Extraction

4. Exploratory Data Analysis

5. Modeling

6. Results

7. Conclusion

8. Questions

Introduction

>>> Project Goal

Distinguish bots from humans

>>> Project Goal

Distinguish bots from humans

* But what about CAPTCHAs?

>>> Project Goal

Distinguish bots from humans

* But what about CAPTCHAs?
* But what about honeypots?

>>> Adversarial Example

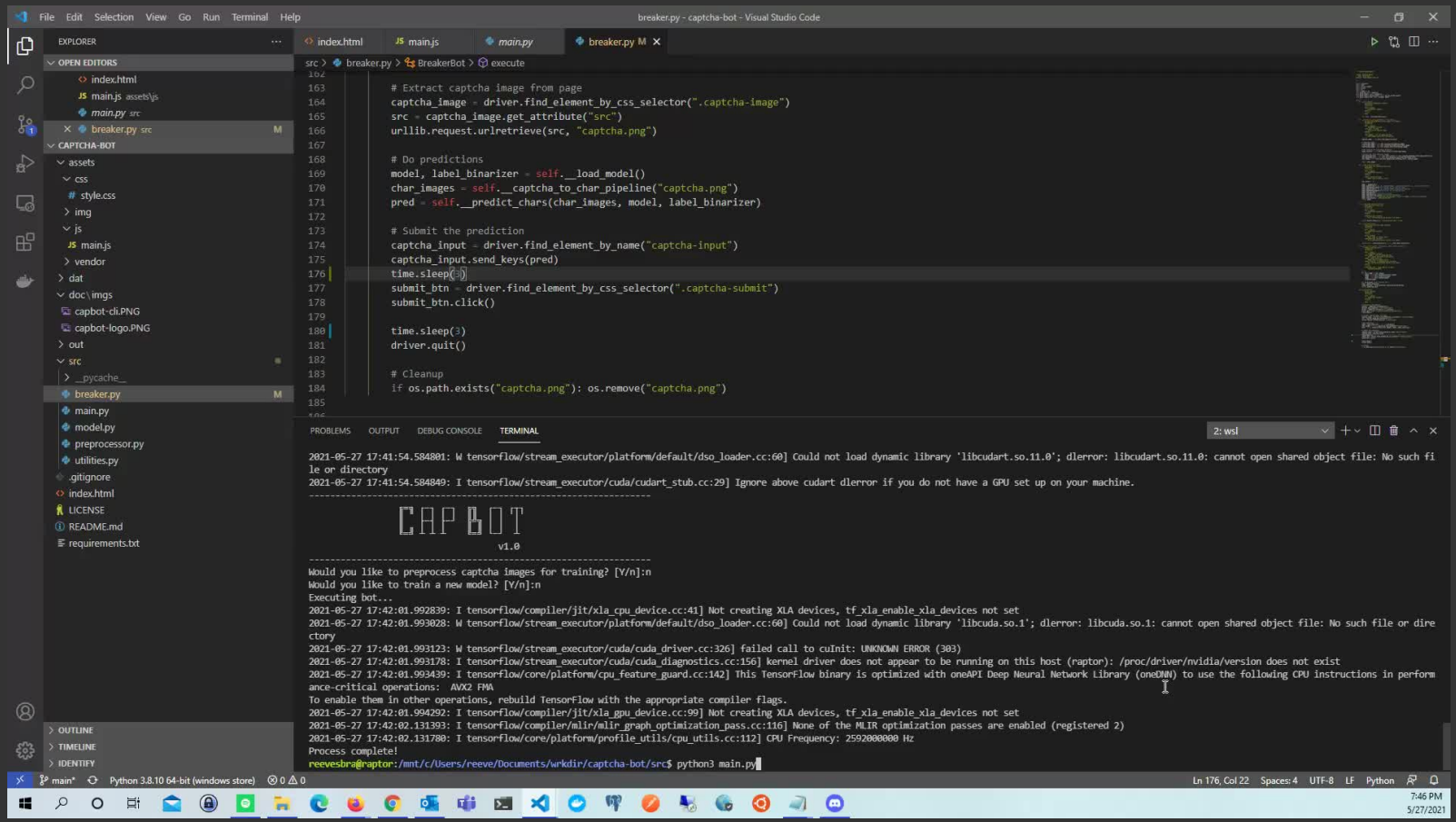

Figure 1: CAPTCHA Breaking Bot[1].

[1] https://github.com/reevesba/capbot

>>> Solution

Biometrics: Intrinsic human characteristics

>>> Solution

Biometrics: Intrinsic human characteristics

1. Physical: Something you are
    * fingerprints, iris patterns, DNA, etc.

>>> Solution

Biometrics: Intrinsic human characteristics

1. Physical: Something you are
    * fingerprints, iris patterns, DNA, etc.
2. Behavioral: Something you do
    * engagement patterns, physical movements, typing patterns, etc.

Data Collection

>>> Keylogging Application

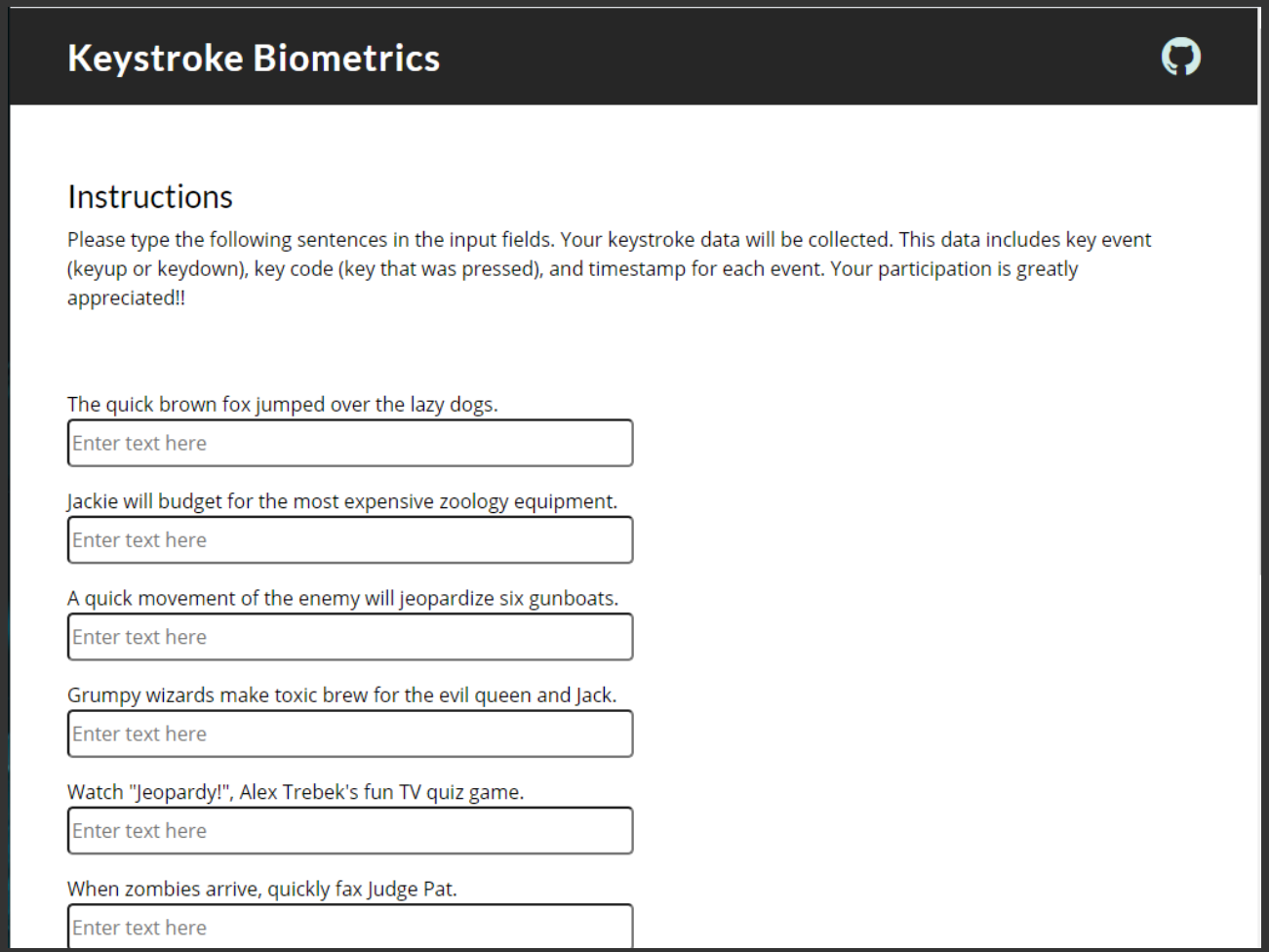

Figure 2: keystrokebiometrics.xyz [2]

[2] https://github.com/reevesba/keystrokebiometrics.xyz

>>> KEYSTROKE_METRICS Table

```
mysql> desc production.KEYSTROKE_METRICS;
+-------------+-------------+------+-----+---------+-------+
| Field       | Type        | Null | Key | Default | Extra |
+-------------+-------------+------+-----+---------+-------+
| uuid        | varchar(50) | YES  |     | NULL    |       |
| sentence_id | int         | YES  |     | NULL    |       |
| key_event   | varchar(10) | YES  |     | NULL    |       |
| key_code    | int         | YES  |     | NULL    |       |
| key_char    | varchar(10) | YES  |     | NULL    |       |
| alt_key     | tinyint(1)  | YES  |     | NULL    |       |
| ctrl_key    | tinyint(1)  | YES  |     | NULL    |       |
| shift_key   | tinyint(1)  | YES  |     | NULL    |       |
| is_bot      | int         | YES  |     | NULL    |       |
| timestamp   | bigint      | YES  |     | NULL    |       |
+-------------+-------------+------+-----+---------+-------+
10 rows in set (0.00 sec)
```

Figure 3: Table Schema.

Feature Extraction

>>> Potential Features

1. Dwell time: Time from keydown to keyup
    * 52+ features (a-zA-Z)

>>> Potential Features

1. Dwell time: Time from keydown to keyup
   * 52+ features (a-zA-Z)
2. Flight time: Time from keydown to keydown
   * 52+ characters gives us $52^2 = 2704$+ features

>>> Potential Features

1. Dwell time: Time from keydown to keyup
   * 52+ features (a-zA-Z)
2. Flight time: Time from keydown to keydown
   * 52+ characters gives us $52^2 = 2704$+ features
3. Characters per minute (typing speed)

>>> Potential Features

1. Dwell time: Time from keydown to keyup
   * 52+ features (a-zA-Z)
2. Flight time: Time from keydown to keydown
   * 52+ characters gives us $52^2 = 2704$+ features
3. Characters per minute (typing speed)

This is too many dimensions, so I reduced them.

>>> Grouping Keys

Four groups: Left hand w/o shift, left hand w/ shift, right hand w/o shift, right hand w/ shift

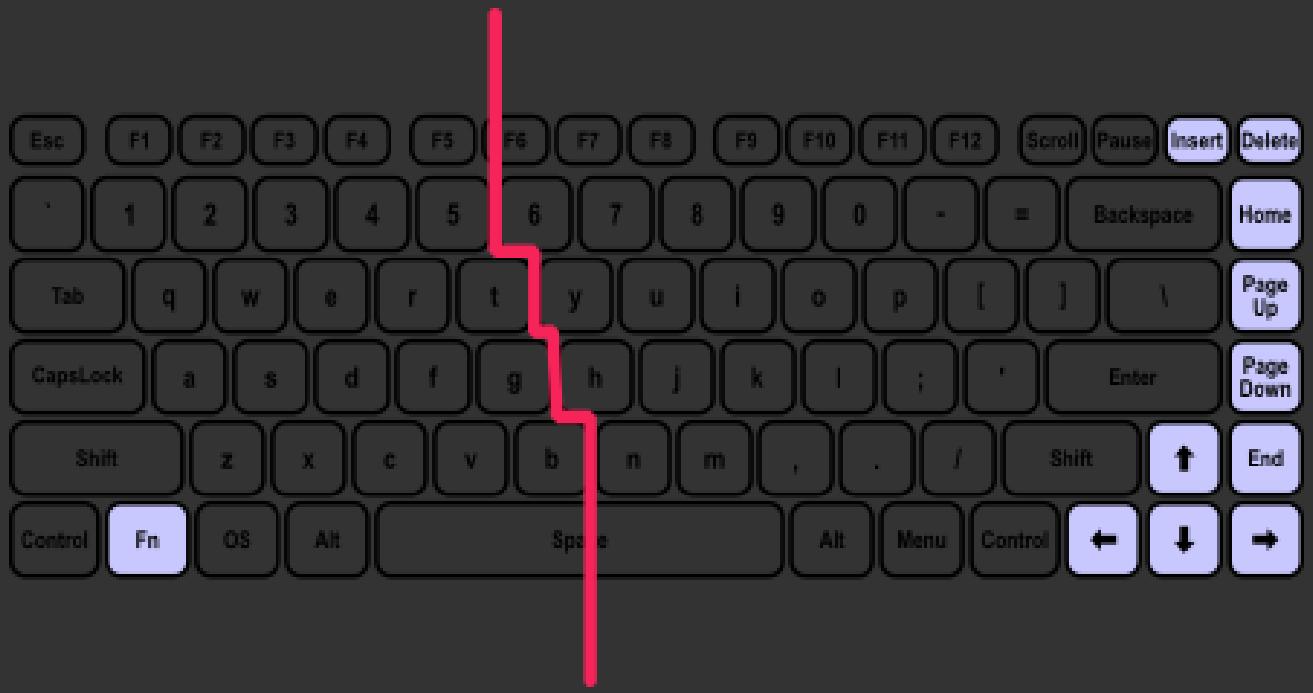

Figure 4: Qwerty Keyboard Split.

>>> Final Features

Category	Feature	Description
Flight Time	lr	Transition from left hand to right hand
	Lr	Transition from left hand + shift to right hand
	rl	Transition from right hand to left hand
	Rl	Transition from right hand + shift to left hand
	ll	Transition from left hand to left hand
	Ll	Transition from left hand + shift to left hand
	rr	Transition from right hand to right hand
	Rr	Transition from right hand + shift to right hand
Dwell Time	l	Hold duration of left hand
	r	Hold duration of right hand
	L	Hold duration of left hand + shift
	R	Hold duration of right hand + shift
	space	Hold duration of space bar
	shift	Hold duration of shift key.
Chars/Min	cpm	Average number of keys pressed in one minute

Table 1: 15 Extracted Features.

>>> Feature Data

	uuid	sentence_id	key_event	key_code	key_char	alt_key	ctrl_key	shift_key	is_bot	timestamp
x_1	1033570f-0efe-4748-a92e...	0	keydown	16	Shift	0	0	1	0	1621639356958
x_2	1033570f-0efe-4748-a92e...	0	keydown	84	T	0	0	1	0	1621639357107
x_2	1033570f-0efe-4748-a92e...	0	keyup	16	Shift	0	0	0	0	1621639357199
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
x_{41700}	01907d83-c61a-4464-8f7d...	9	keyup	190	.	0	0	0	0	1621910136024

Table 2: Raw Data Peek.

	lr	Lr	rl	Rl	ll	Ll	rr	Rr	l	r	L	R	space	shift	cpm	is_bot
x_1	145.58	90.0	132.33	0	135.0	0	86.5	0	1.75	1.41	1.0	0	1.75	2.0	19.37	1
x_2	112.57	0.0	125.54	76.0	130.54	0	103.3	0	1.48	1.5	0.0	1.0	1.0	2.0	20.27	1
x_2	105.4	0.0	117.36	0	158.5	209.0	122.11	0	1.68	1.46	2.0	0	1.78	2.0	19.62	1
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
x_{365}	208.38	345.0	125870.2	137.0	3480.3	0	235.5	184.0	151.0	155.36	178.0	95.0	149.67	229.67	0.09	0

Table 3: Biometrics Data Peek.

Exploratory Data Analysis

>>> Outlier Detection

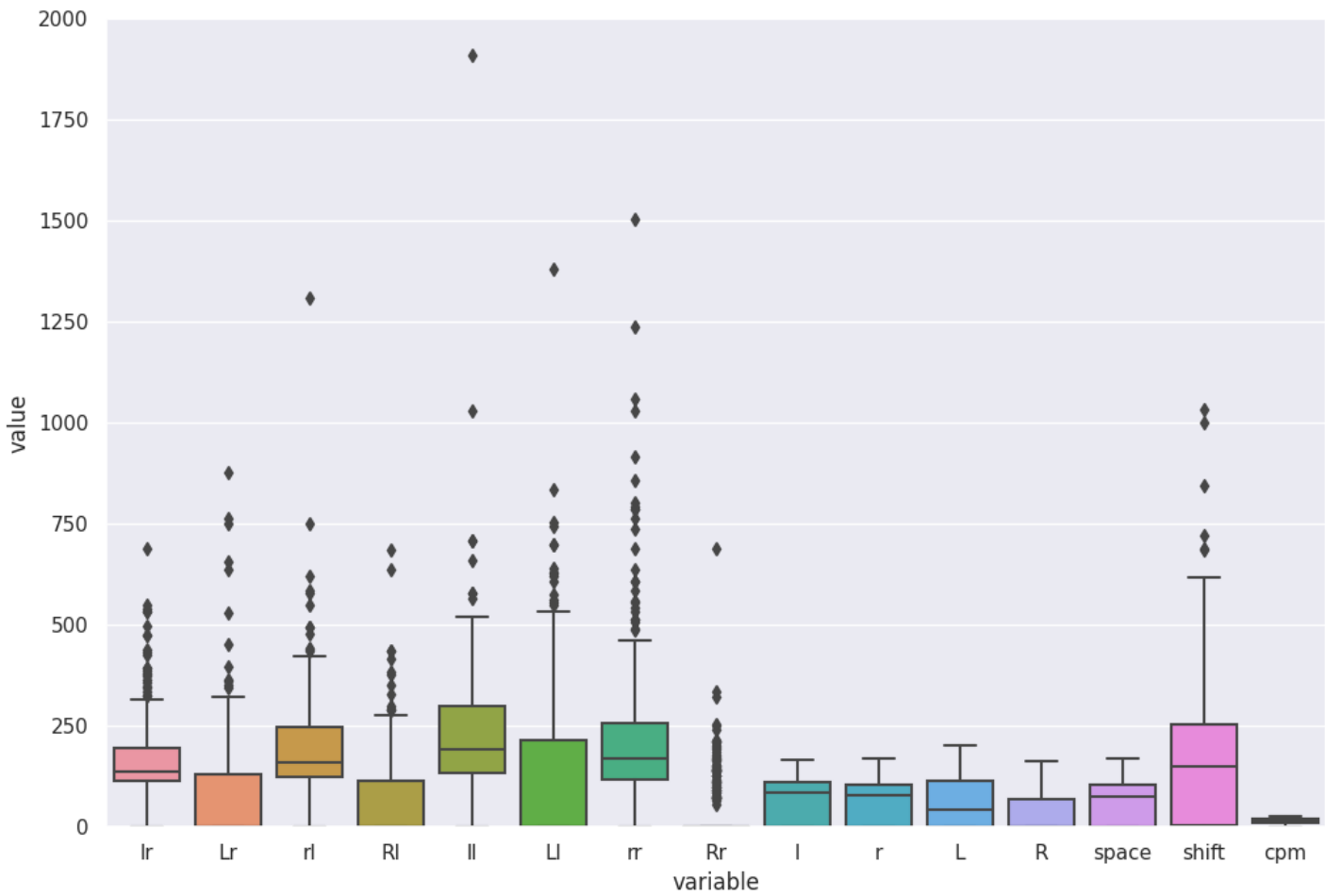

Figure 5: Feature Boxplots.

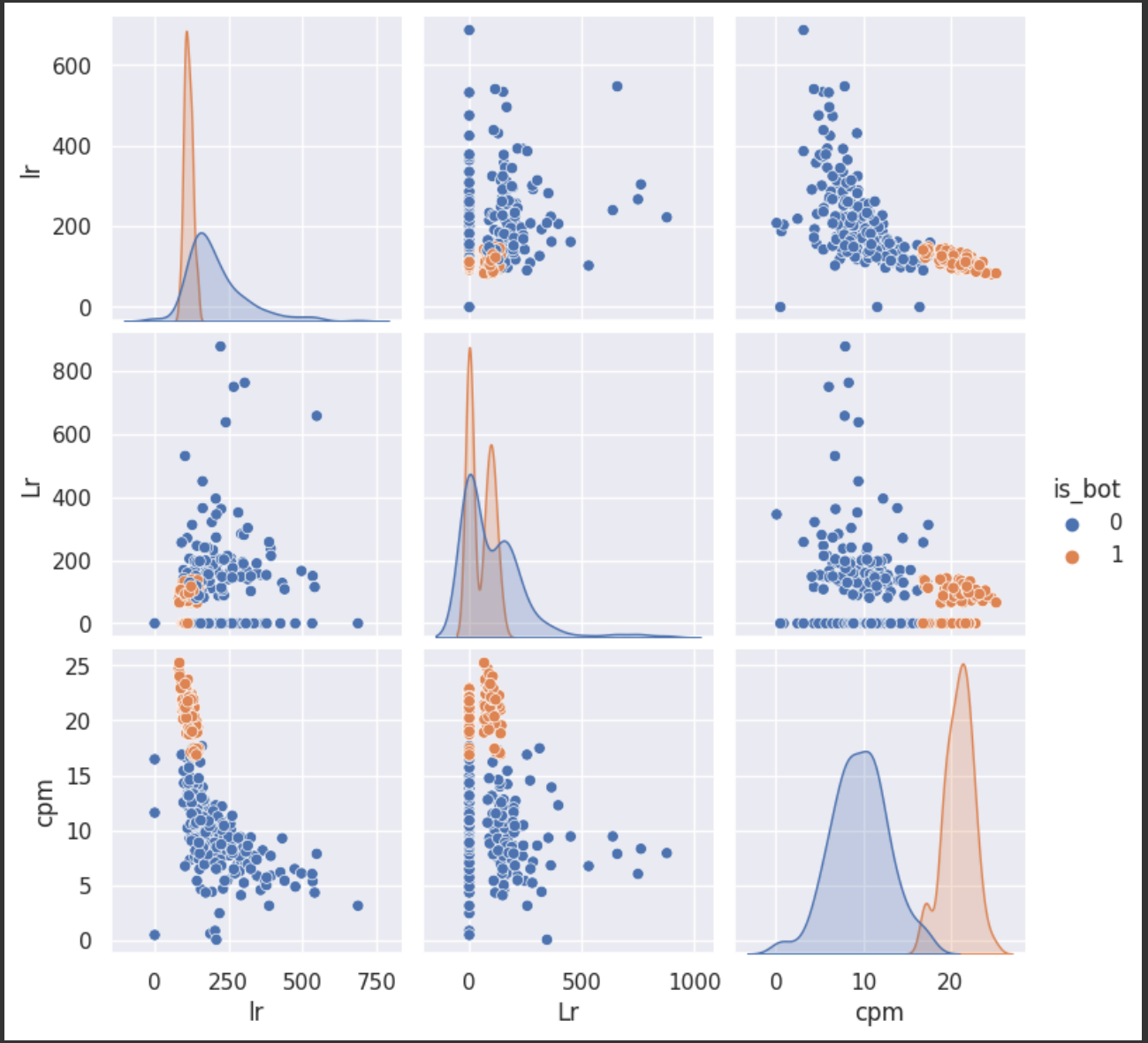

Figure 6: Feature Pairplot.

>>> Sample Statistics

	cpm						
	count	mean	std	min	25%	50%	75%	max
215.0	9.567072	3.265893	0.085485	7.483683	9.420892	11.582152	17.649531
150.0	20.871340	1.657686	16.797467	19.863755	21.005201	21.989776	25.219298

Figure 7: cpm Describe.

	lr							
	count	mean	std	min	25%	50%	75%	max
is_bot								
0	215.0	206.766273	101.168686	0.0	142.557692	176.857143	238.311111	686.728814
1	150.0	112.627513	14.049436	81.5	102.239744	111.923077	122.537500	148.750000

[2 rows x 120 columns]

Figure 8: lr Describe.

Modeling

>>> Algorithm Selection

For each algorithm, 5-fold cross-validation was used for evaluation.

1. Naive Bayes
2. Logistic Regression
3. K-Nearest Neighbors
4. Support Vector Machine
5. Decision Tree
6. Decision Tree w/ Bagging
7. Decision Tree w/ Boosting
8. Random Forest
9. Voting classifier w/ Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machine
10. Multilayer Perceptron

>>> Grid Search

Each of the algorithms were trained with sklearn's default parameters. After evaluating each model, a grid search was performed on the top two algorithms.

1. Random Forest
    * n_estimators: [10, 100, 1000]
    * max_depth: [5, 10, 15]
    * min_samples_leaf: [1, 2, 5, 10]
    * max_leaf_nodes: [2, 5, 10]
2. K-Nearest Neighbors
    * n_neighbors: [5, 10, 15]
    * weights: [uniform, distance]
    * metric: [euclidean, manhattan]

Results

>>> Initial Results

Algorithm	Score
Naive Bayes	0.9726
Logistic Regression	0.9726
K-Nearest Neighbors	0.9945
Support Vector Machine	0.9726
Decision Tree	0.9726
Decision Tree w/ Bagging	0.9534
Decision Tree w/ Boosting	0.9726
Random Forest	0.9836
Voting classifier	0.9753
Multilayer Perceptron	0.9781

Table 4: Estimator Scores.

>>> Grid Search Results

Best Estimator

- Classifier: Random Forest
- n_estimators: 100
- max_depth: 5
- min_samples_leaf: 1
- max_leaf_nodes: 5

>>> Test Data Evaluation

	0	1
0	53	0
1	0	39

Table 5: Confusion Matrix.

	Precision	Recall	F1 Score	Support
0	1.00	1.00	1.00	53
1	1.00	1.00	1.00	39
Accuracy			1.00	92
Macro Avg	1.00	1.00	1.00	92
Weighted Avg	1.00	1.00	1.00	92

Table 6: Classification Report.

Conclusion

>>> Closing Thoughts

* Bot detection is very easy using keystroke biometrics

>>> Closing Thoughts

* Bot detection is very easy using keystroke biometrics
* Can be extended to user authentication

>>> Closing Thoughts

* Bot detection is very easy using keystroke biometrics
* Can be extended to user authentication
* One problem is that user's behaviors may change

>>> Closing Thoughts

* Bot detection is very easy using keystroke biometrics
* Can be extended to user authentication
* One problem is that user's behaviors may change
* Another problem is the legality of keylogging software (can be considered wire-tapping)

>>> Questions?

bradley.reeves@cwu.edu