Anomaly Detection in Host Log Files

Kate Stadelman

science. engineering. analytics. insights.

DATA

Research Question

Anomaly Detection in Host Log Files

Can we detect suspicious user activity hidden in massive computer log files?

Unified Network & Network Data Set

"A subset of network and computer (host) events collected from the Los Alamos National Laboratory enterprise network over the course of approximately 90 days."[1]

Details

- Microsoft Windows Computers
- Logs Deidentified for Security
- Day One: 55.6M Events

Hacking 101

Low-Level User
Administrator

1. Gain Access to a Computer with a Low-Level User
2. Escalate Privileges (to Administrator)
3. Run Harmful Code
   - Hijack Processes
   - Steal / Destroy Data
   - Breach Other Computers on the Network

Connect Logs By

Event Time

Computer (Host) Name

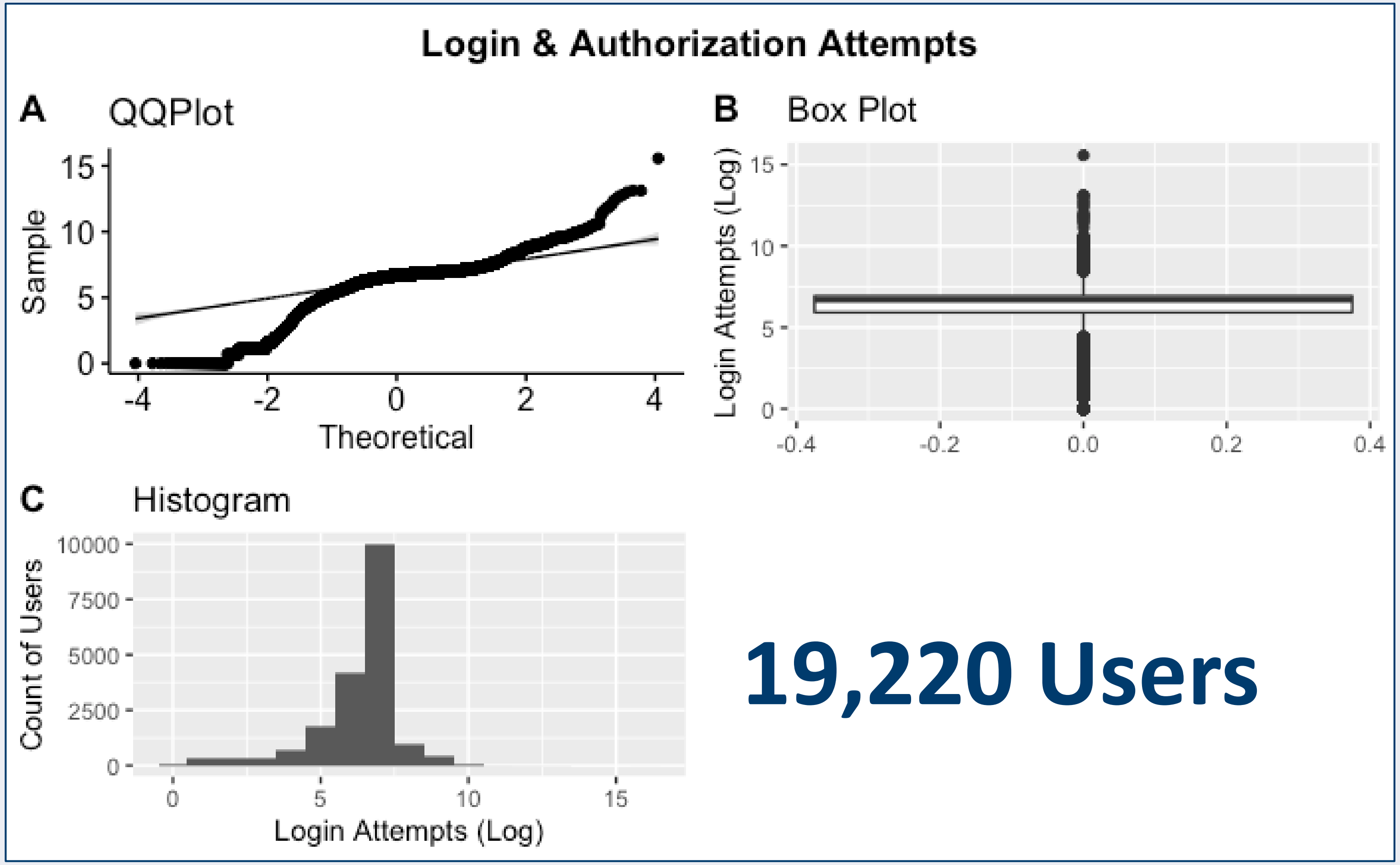
User Name

Logon ID

Computer (Host) Events

- Windows Authentications (Kerberos)
- Logins Using Explicit Credentials
- Login Failures (with Reason Codes)
- Special Privileges Assigned to Login
- Process Started / Stopped
- Computer Locked / Unlocked
- Screen Saver Started / Stopped

Log Aggregation

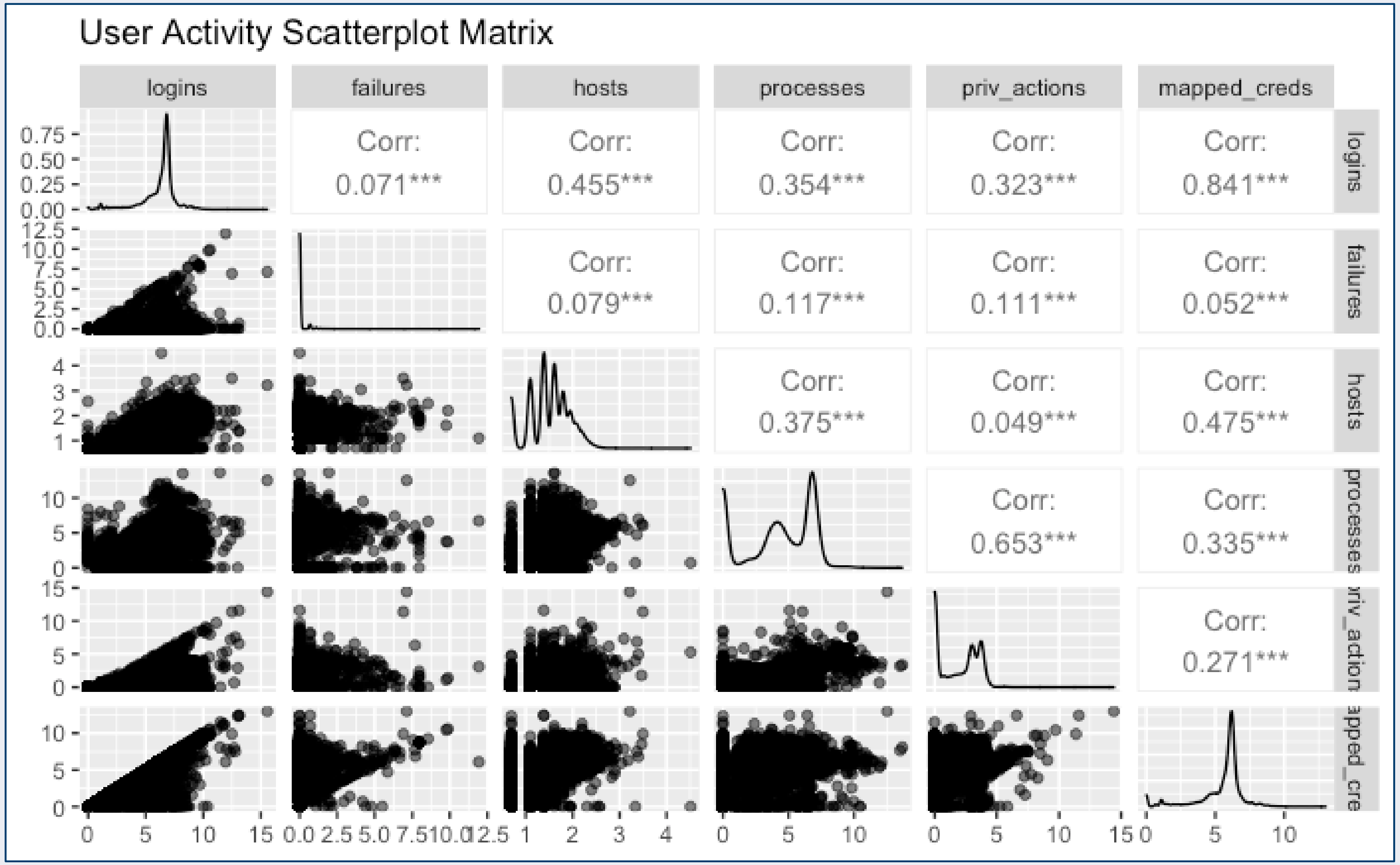


19,220 Users

Counts by User

- Login Attempts
- Login Failures
- Hosts Accessed
- Privileged Actions
- Processes Started
- Mapped Credentials

Variable Check



19,220 Users

Isolation Forest

- First Presented in 2008
- Developed for Anomaly Detection
- Unsupervised Tree Ensemble Method
- Extension of Random Forest
- Scales Well for Large Data Sets

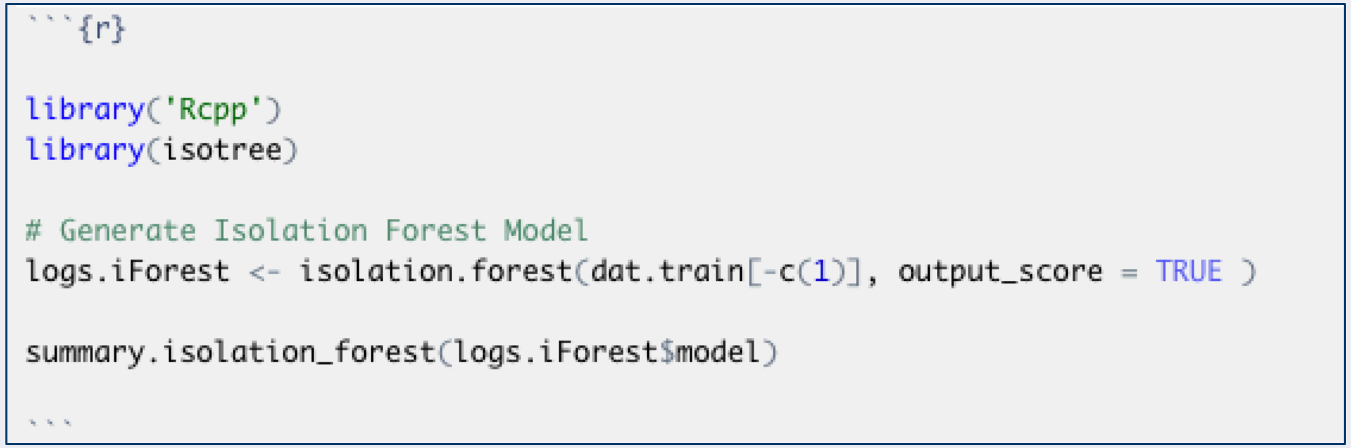
Which One is Different?

Data Requirements

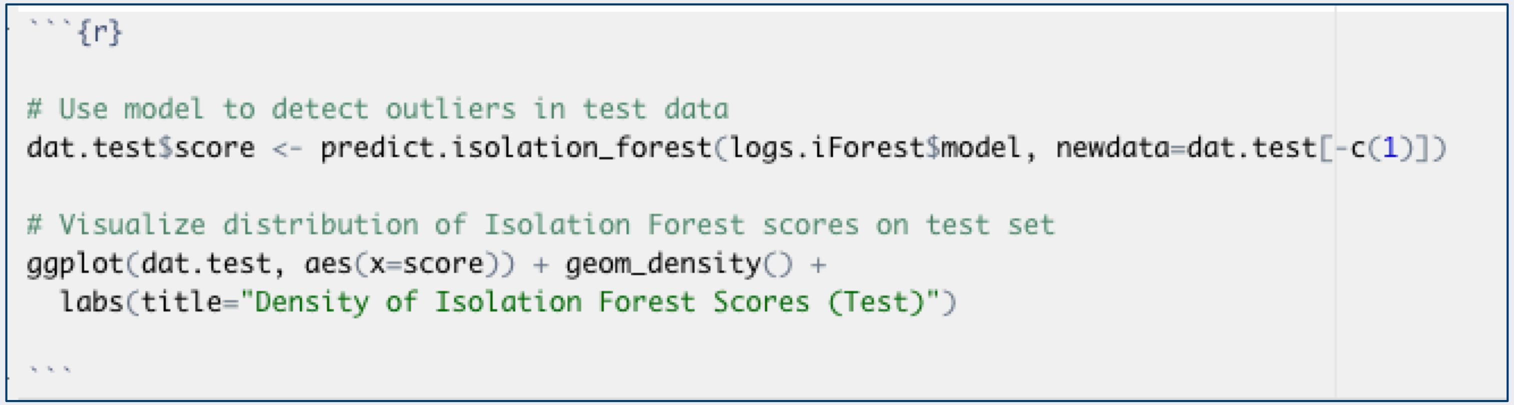
Bad Data? Isolation Forest STILL WORKS

- High-Dimension Problems
- Data is Not Well-Behaved
- Large Number of Irrelevant Attributes
- Small Training Sets
- Training Sets Without Anomalies

isotree Package

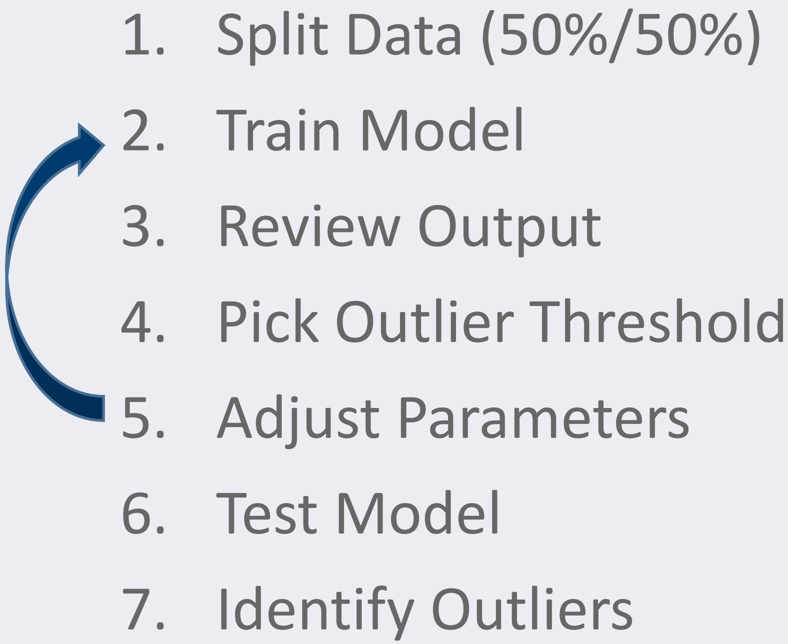


~~~
```{r}

library('Rcpp')
library(isotree)

# Generate Isolation Forest Model
logs.iForest <- isolation.forest(dat.train[-c(1)], output_score = TRUE )

summary.isolation_forest(logs.iForest$model)

```
~~~



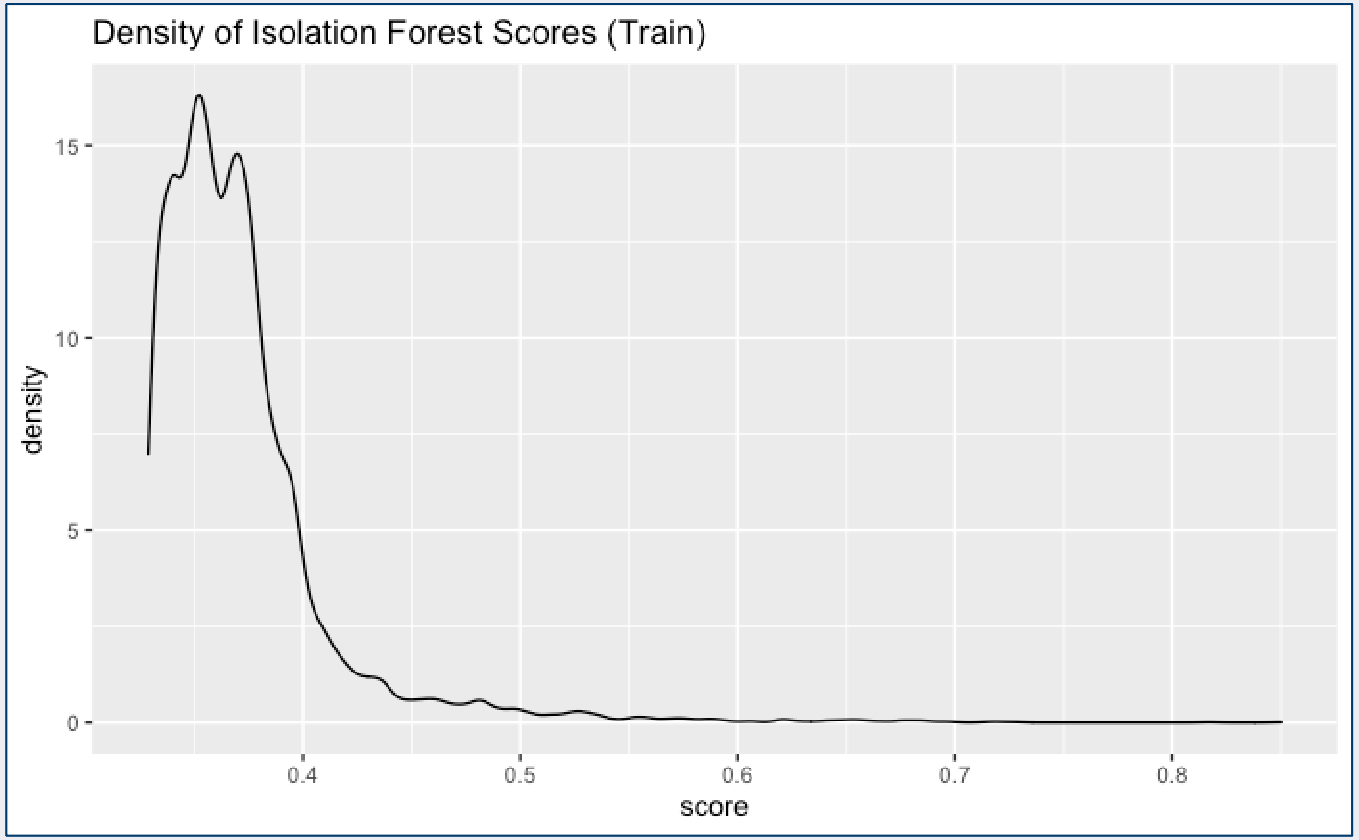
~~~
```{r}

# Use model to detect outliers in test data
dat.test$score <- predict.isolation_forest(logs.iForest$model, newdata=dat.test[-c(1)])

# Visualize distribution of Isolation Forest scores on test set
ggplot(dat.test, aes(x=score)) + geom_density() +
  labs(title="Density of Isolation Forest Scores (Test)")

```
~~~

Process



1. Split Data (50%/50%)
2. Train Model
3. Review Output
4. Pick Outlier Threshold
5. Adjust Parameters
6. Test Model
7. Identify Outliers

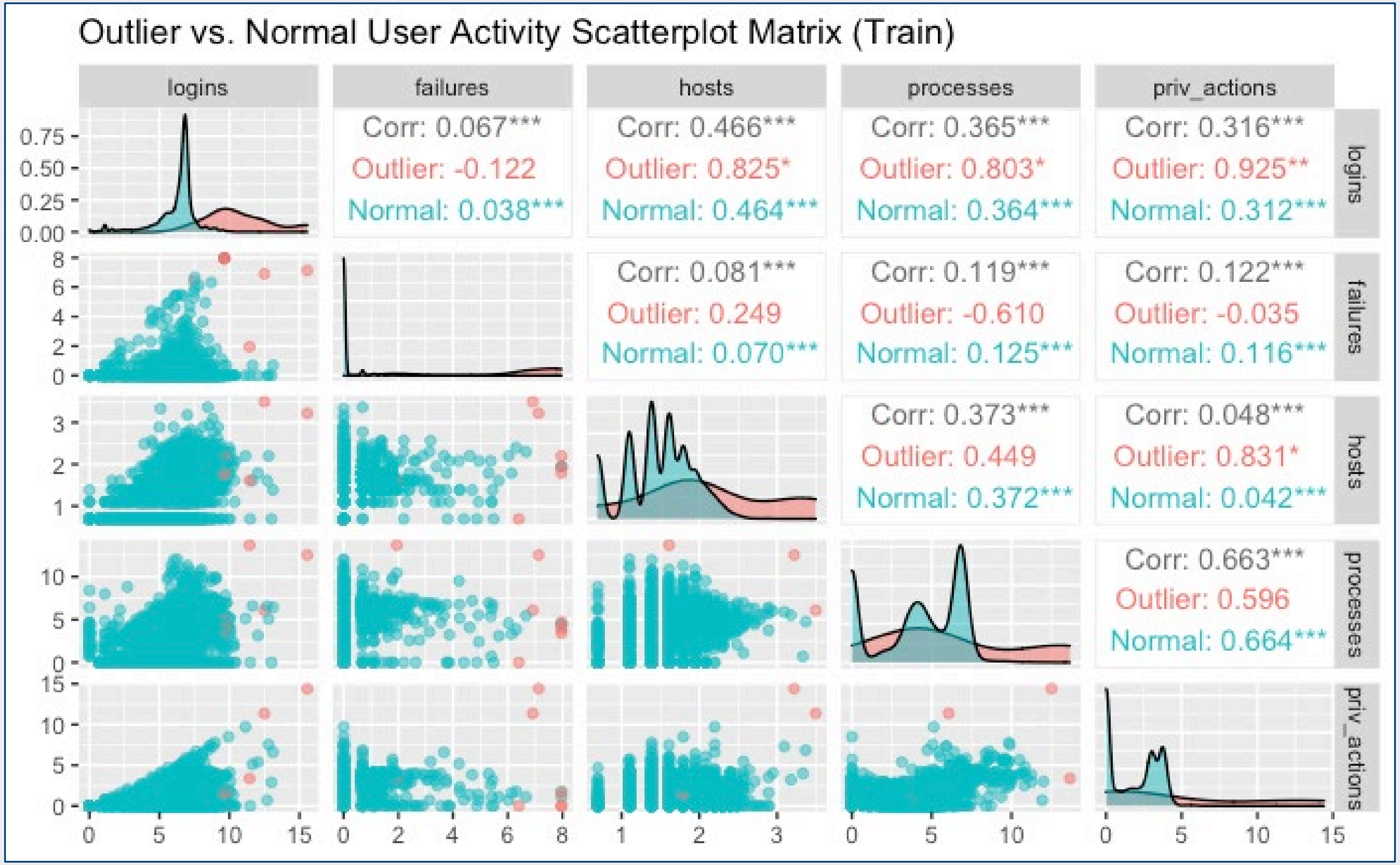
Model Output


Density of Isolation Forest Scores (Train)

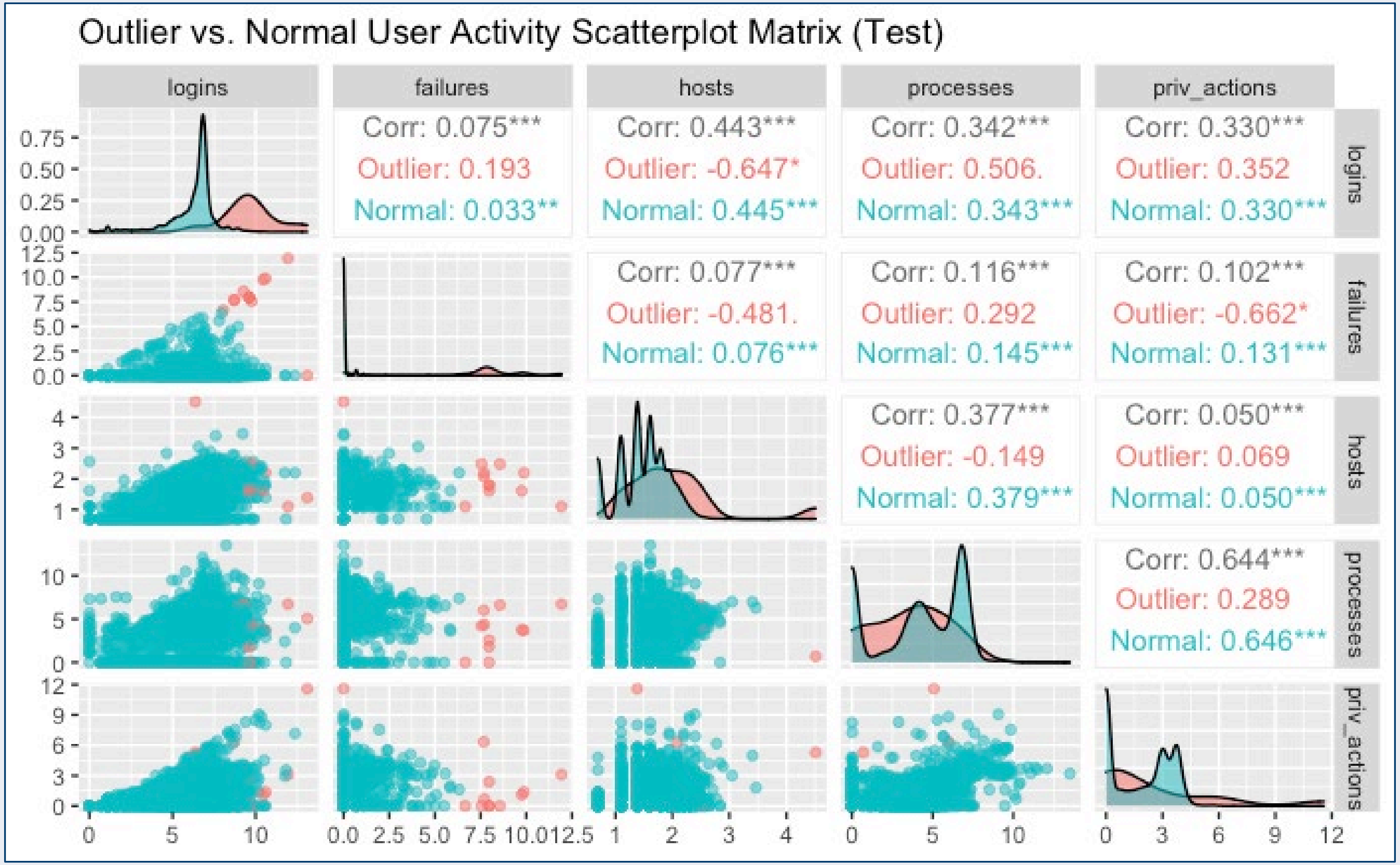
Anomaly Scores

- Generated for Each Observation
- Derived From Average Path Length
- Closer to 0.5 → Normal
- Closer to 1 → Outlier
- Set Outlier Threshold for Analysis (0.7)

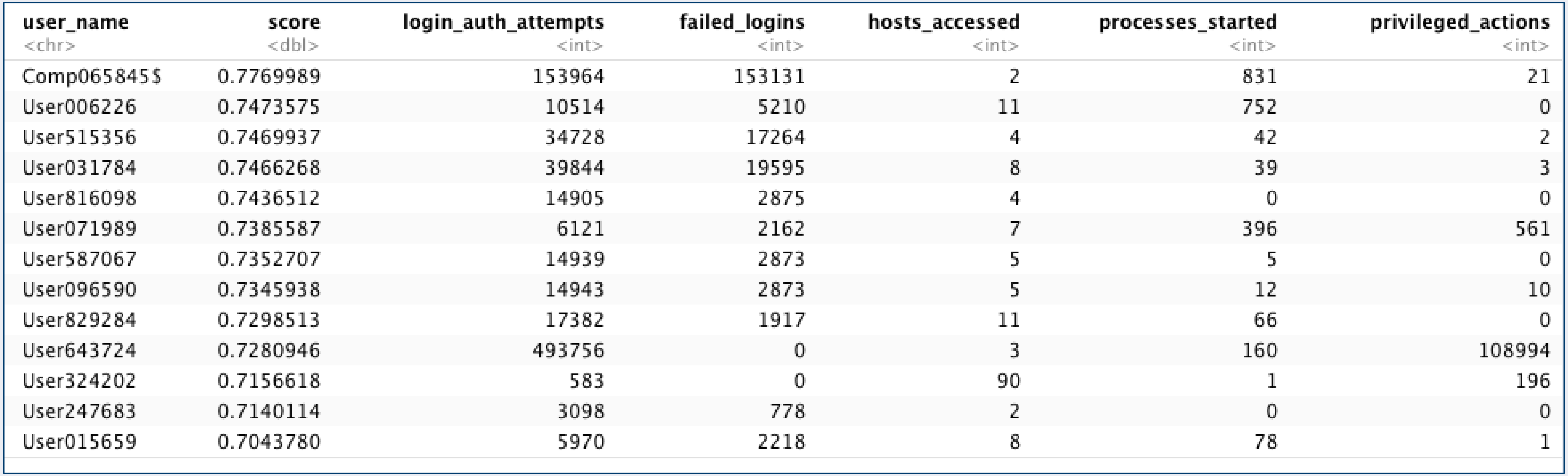
Training Results



Test Results



Anomalous Users



user_name	score	login_auth_attempts	failed_logins	hosts_accessed	processes_started	privileged_actions
<chr>	<dbl>	<int>	<int>	<int>	<int>	<int>
Comp065845$	0.7769989	153964	153131	2	831	21
User006226	0.7473575	10514	5210	11	752	0
User515356	0.7469937	34728	17264	4	42	2
User031784	0.7466268	39844	19595	8	39	3
User816098	0.7436512	14905	2875	4	0	0
User071989	0.7385587	6121	2162	7	396	561
User587067	0.7352707	14939	2873	5	5	0
User096590	0.7345938	14943	2873	5	12	10
User829284	0.7298513	17382	1917	11	66	0
User643724	0.7280946	493756	0	3	160	108994
User324202	0.7156618	583	0	90	1	196
User247683	0.7140114	3098	778	2	0	0
User015659	0.7043780	5970	2218	8	78	1

Discussion

Q & A

science.engineering.analytics.insights.

DATA

References

Turcotte M., Kent A., & Hash C. (2018, November). Unified Host and Network Data Set. *Data Science for Cyber-Security*, (1–22). https://www.worldscientific.com/doi/abs/10.1142/9781786345646_001

Liu F. T., Ting K. M., & Zhou, Z.-H. (2008). *Isolation Forest*. Proceedings of the 8th IEEE International Conference on Data Mining (ICDM’08), Pisa, Italy. https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf

Zhu, A. & Suresh, S. (2019, December 13). *Isolation Forest for Data Mining*. Medium. https://medium.com/@siddharth.suresh92/isolation-forest-for-data-mining-a2c44a26d646

Young, A. (2020, November 13). *Isolation Forest is the best Anomaly Detection Algorithm for Big Data Right Now*. Towards Data Science. https://towardsdatascience.com/isolation-forest-is-the-best-anomaly-detection-algorithm-for-big-data-right-now-e1a18ec0f94f

science.engineering.analytics.insights. DATA