Splunk4Ninjas - Machine Learning

Lab Guide

Overview

This lab guide contains the hands-on exercises for the Splunk4Ninjas - Machine Learning workshop. Before proceeding with these exercises, please ensure that you have a copy of the workshop slide deck, which will help to put into context the tasks you are carrying out.

Download the workshop slide deck: https://splk.it/S4N-ML-Attendee

Prerequisites

Splunk.com account

In order to complete these exercises, you will need your own Splunk instance. Splunk's hands-on workshops are delivered via the Splunk Show portal and you will need a splunk.com account in order to access this.

If you don't already have a Splunk.com account, please create one here before proceeding with the rest of the workshop.

Splunk Knowledge

A working knowledge of Splunk is assumed for this workshop.

Table of Contents

Access Your Lab Environment

Description

You'll need a Splunk instance to do these hands-on exercises – time to get one!

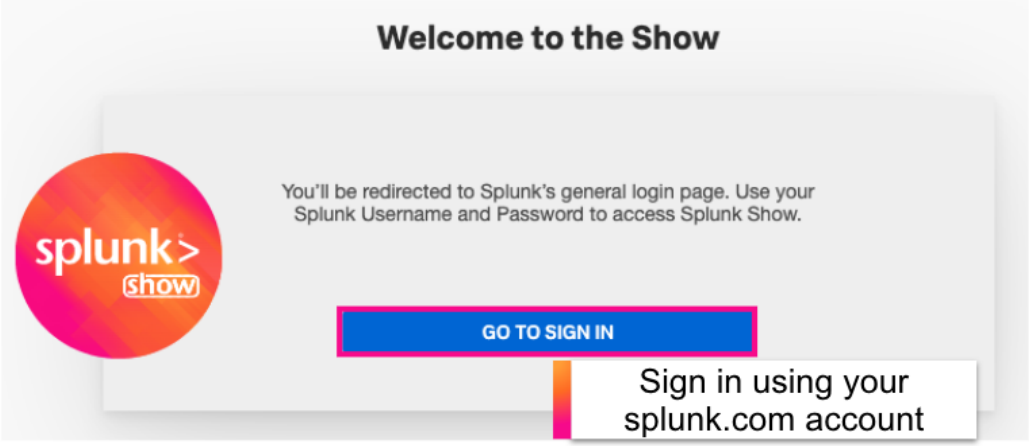
In this exercise, you will create your own Splunk Enterprise instance using our Splunk Show portal.

Already been given your Splunk instance details?

If your workshop host has already provided you with your instance URL and login details then you do not need to follow the instructions in this first section - you can skip straight to Challenge 1!

Steps

1. Browse to https://show.splunk.com and log in using your **Splunk.com account**.



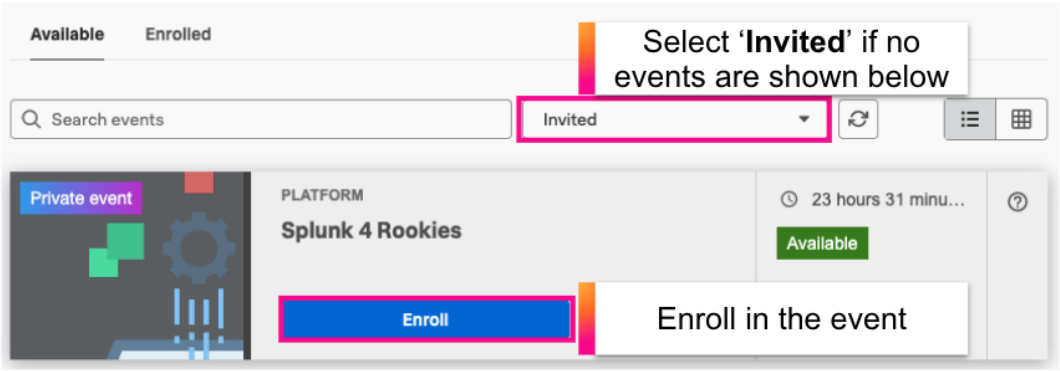
🔑 **Don't have a Splunk.com Account?**

To access our hands-on workshop events you will need a Splunk.com account. If you don't already have a Splunk.com account, don't worry - it only takes a few minutes to create one! Please create one here.

---

2. Once logged in to Splunk Show you will see the event page for the event that you have been invited to. If no events are listed, try selecting '**Invited**' from the dropdown list.

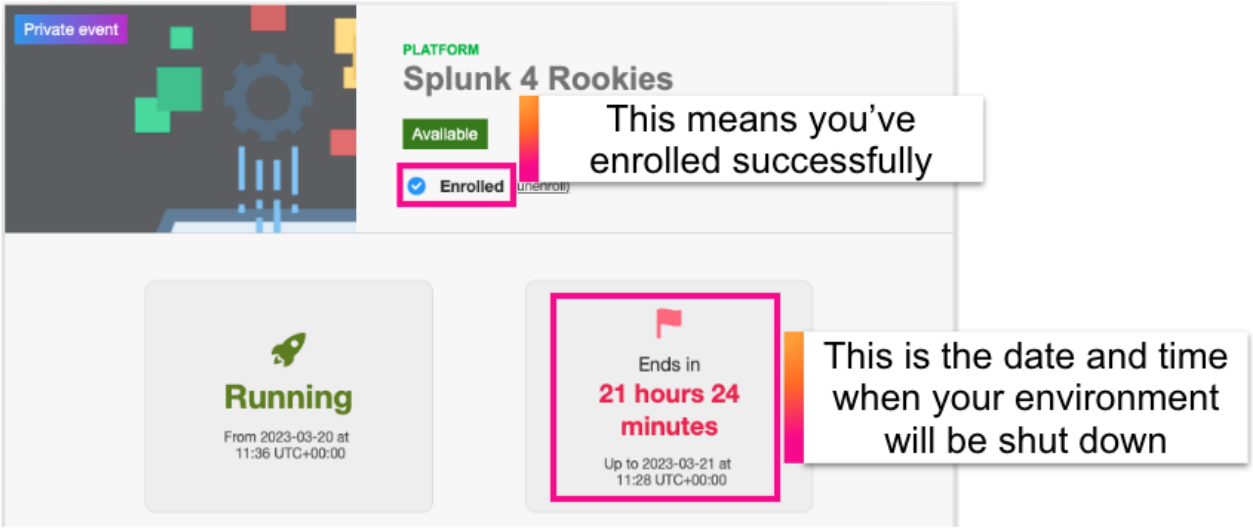
Click on **Enroll** to join the event.



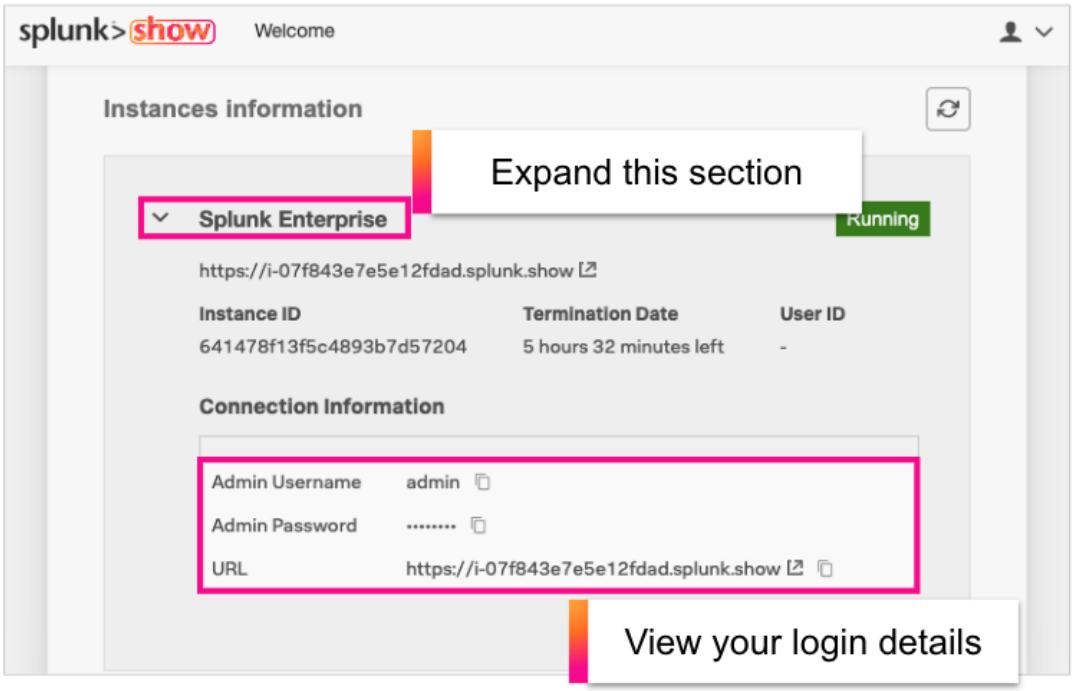
The page will refresh and the event will now display 'Enrolled'.

🕐 Lab environment expiration

All Splunk environments that are part of this workshop event will automatically be shut down at the date and time specified on this screen so feel free to continue to play around with your lab environment until then!

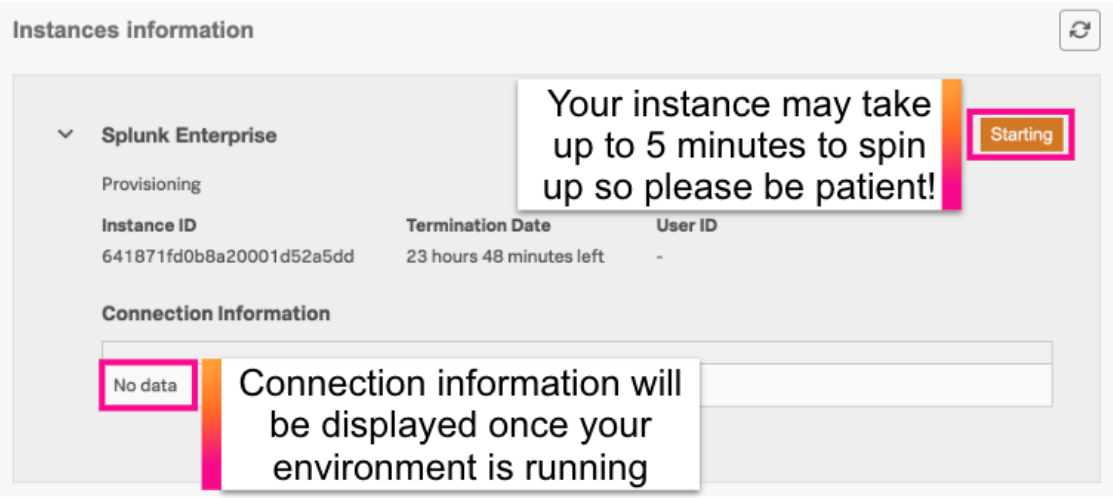


3. Scroll down the page to the **Instances Information** section and expand out the 'Splunk Enterprise' section to locate the user credentials and link to your lab environment.



ℹ️ No connection information shown?

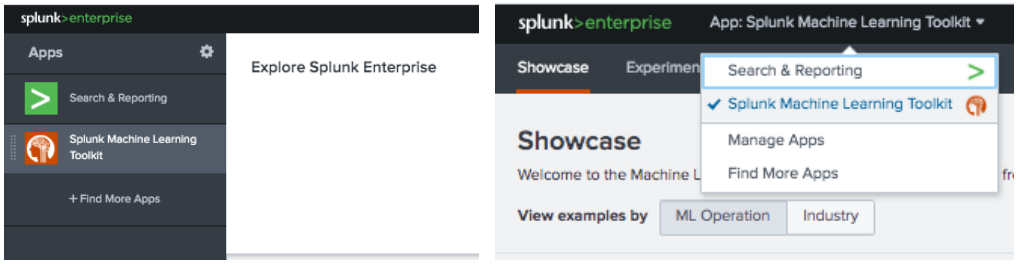
If you don't see any connection information displayed yet it means that your lab environment is currently starting up. Please try refreshing this view in a few minutes.



Challenge 1 – Create a Sample Dataset

Steps

1. Select the Splunk Machine Learning Toolkit (MLTK) app (all work completed in this session should be done in this app):



2. Open Search in MLTK and examine the track_day.csv lookup:

```
| inputlookup track_day.csv
```

3. Rename fields with x_ or y_ prefix:

```
| inputlookup track_day.csv
| rename * as x_*, x_vehicleType as y_vehicleType
```

4. Create a sample of 1000 results per y_vehicleType (target variable) which provides the same random set for everyone who uses the data set:

```
| inputlookup track_day.csv
| rename * as x_*, x_vehicleType as y_vehicleType
| sample 1000 by y_vehicleType seed=123
```

5. Use this sample of 6000 events to create a new dataset called “mytrackdata.csv”

```
| inputlookup track_day.csv
| rename * as x_*, x_vehicleType as y_vehicleType
| sample 1000 by y_vehicleType seed=123
| outputlookup mytrackdata.csv
```

6. Confirm your new lookup exists in Settings > Lookups > Lookup Table Files or by running a search:

```
| inputlookup mytrackdata.csv
```

7. Explore the new fields within your lookup:

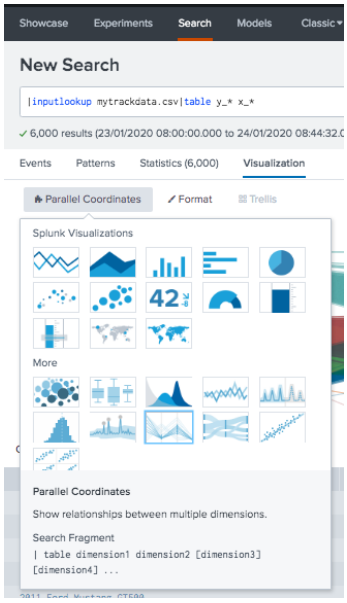
```
| inputlookup mytrackdata.csv
| fieldsummary
```

Create a parallel coordinates chart

8. Create a parallel coordinates chart to help understand the different field values that each vehicle type is represented by:

```
| inputlookup mytrackdata.csv
| table  y_* x_*
```

Select the “Visualizations” tab and ensure that “Parallel Coordinates” is selected as the visualization type.



9. Save your visualization to a new dashboard:

Dashboard Title: Track Day Exploration
Panel Title: Parallel Coordinates by Vehicle

Create a histogram

10. Use a macro (you must bookend the macro and the variable values with the backtick symbol `) to create a histogram by battery voltage:

```
| inputlookup mytrackdata.csv
| `histogram(x_batteryVoltage, 1000)`
```

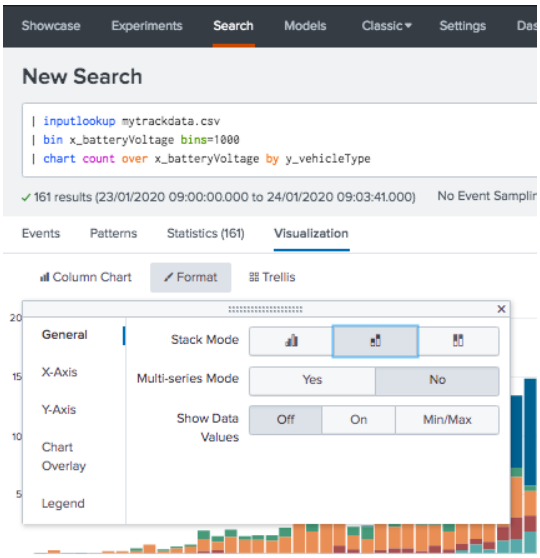
Select the Visualization tab and choose the “Histogram” visualization and save to a dashboard:

Dashboard Title: Track Day Exploration
Panel Title: Battery Voltage Histogram

11. Create a column chart (that resembles a histogram) of battery voltage “colored” by vehicle type:

```
| inputlookup mytrackdata.csv
| bin x_batteryVoltage bins=1000
| chart count over x_batteryVoltage by y_vehicleType
```

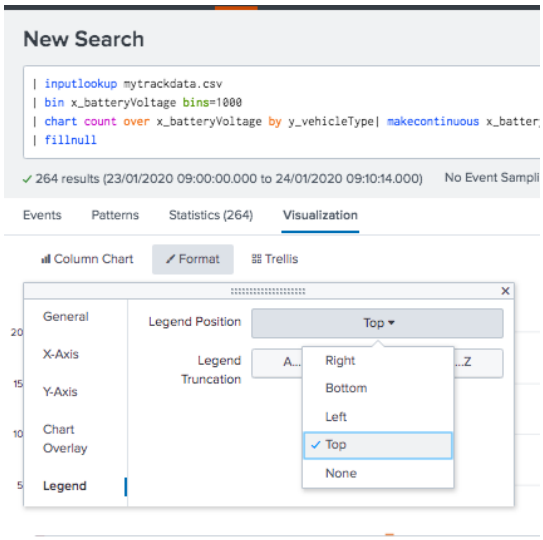
12. Select Column Chart visualization and “Stack” the columns.



13. Add to the search to give the column chart histogram features:

```
| inputlookup mytrackdata.csv
| bin x_batteryVoltage bins=1000
| chart count over x_batteryVoltage by y_vehicleType
| makecontinuous x_batteryVoltage
| fillnull
```

14. Move legend to the top of the visualization and save to dashboard:



Dashboard Title: Track Day Exploration
Panel Title: Battery Voltage by Vehicle Type Histogram

Create a boxplot

15. Use an existing macro to create a box plot chart on all the numeric fields you are using with scaled values:

```
| inputlookup mytrackdata.csv
| fit StandardScaler x_*
```

16. Table your newly created scaled values and create a “Boxplot Chart” visualisation and save to a dashboard:

```
| inputlookup mytrackdata.csv
| fit StandardScaler x_*
| table SS_*
| `boxplot`
```

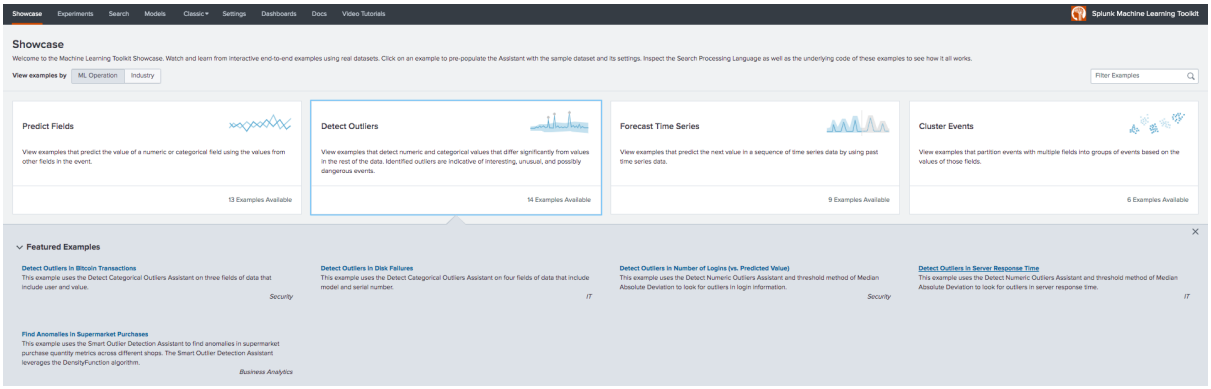
Dashboard Title: Track Day Exploration
Panel Title: Parameter Range Boxplot

Challenge 2 – Detect Outliers In Your Dataset

Steps

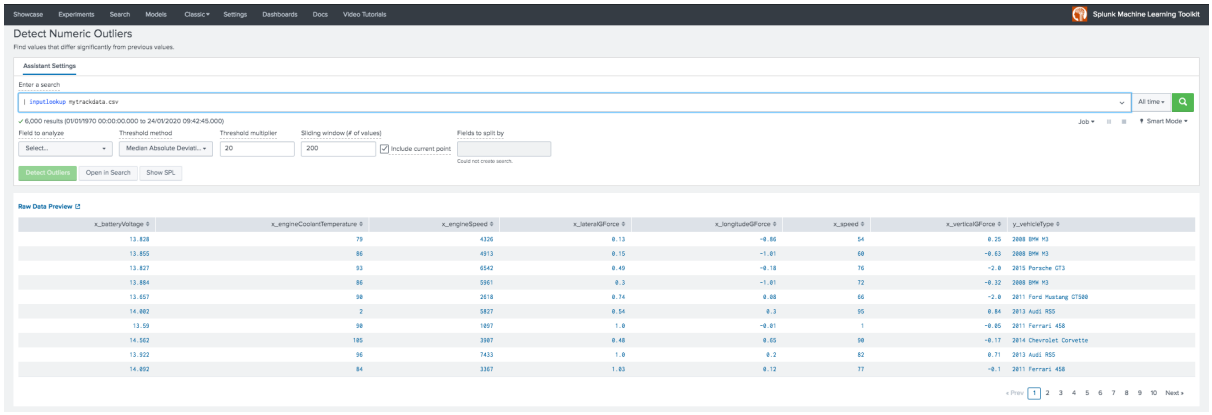
1. Explore the MLTK Showcase:

   Open Showcase > Detect Outliers > Detect Outliers in Server Response Time



2. Adapt the showcase to your dataset by searching mytrackdata.csv:

   ```
   | inputlookup mytrackdata.csv
   ```



3. Analyze standard deviation of x_speed by using the dropdown menus:

   Field to analyze : x_speed
   Threshold method: Standard Deviation
   Threshold multiplier: 10 then 5 then 3 (click “Detect Outliers” with various values until you see a good fit, i.e., a reasonable number of outliers)

4. Save the visualization to a dashboard panel:

   Dashboard Title: ML Experiments
   Panel Title: Speed Outlier Detection

5. Create your own experiment:

   Experiments > Smart Outlier Detection

   Experiment Title: outliertrackday

6. Use the mytrackdata.csv dataset:

   ```
   | inputlookup mytrackdata.csv
   ```

7. Detect outliers for x_verticalGForce:

   Click on “Learn”

   Field to analyze: x_verticalGForce
   Split by fields: y_vehicleType
   Distribution type: Normal

   Click “Detect Outliers”

   Click “Split Charts”

   Click on “Review”

8. Save experiment and publish model:

   Click “Save and Next”:

   Experiment Title: outliertrackday

   Click “Save”

   Click “Publish Outlier Models”

   New Main Model Title: outliervehiclegforce

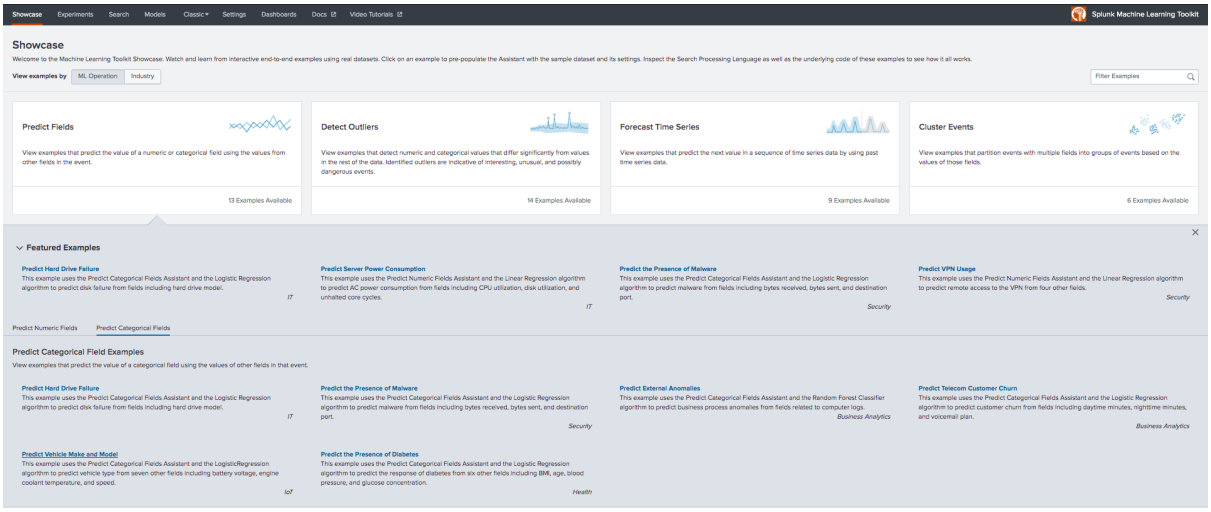
9. Click “Submit” and then “OK”

Challenge 3 – Use A Classifier Model To Predict Vehicle Types From Sensor Data

Steps

1. Explore the MLTK Showcase:

   Open Showcase > Predict Fields > Predict Categorical Fields > Predict Vehicle Make and Model



2. Examine the experiment and consider a few questions:

   What dataset is it using? Is it using any preprocessing steps? What algorithm is it using? What is the split between training and test? What are the best and worst predictions in the Classification Results (Confusion Matrix)?

3. Create your own experiment:

   Experiments > Create New Experiment

   Experiment Type: Predict Categorical Fields
   Experiment Title: Vehicle Category Prediction

   Click “Create”

4. Set the parameters for your experiment:

   Enter a search:
   ```
   | inputlookup mytrackdata.csv
   ```

   Algorithm: LogisticRegression
   Field to predict: y_vehicleType
   Fields to use for predicting: “Select All”
   Split for training/test: 70/30

   Click “Fit to Model”

5. Examine the results of this experiment and then run step 37 with different algorithms:

   RandomForestClassifier, SVM

6. Select the appropriate experiment:

   Open the Experiment History tab

   Expand the most appropriate model (hint, it's probably RandomForestClassifier)

   Select "Load Settings" in the Actions column

   Click "Fit Model"

7. Add a preprocessing step:

   Preprocess method: StandardScaler

   Select the fields to preprocess: [Select all fields except y_vehicleType]

8. Click "Apply"

   Field to predict: y_vehicleType
   Fields to use for predicting: [Select all field with "SS_x" prefix]

   Click "Fit Model"

9. Save the experiment:

   Click "Save"

   Experiment Title: Vehicle Category Prediction

   Click "Save"

   Click "Go to Listings Page"

10. Publish the experiment:

    Click "Publish" for Vehicle Category Prediction

    New Main Model Title: carclassifier
    Destination App: Splunk Machine Learning Toolkit

    Click "Submit"

11. Click "OK"

Challenge 4 – Build A Clustering Model

Steps

1. Explore the MLTK Showcase:

   Click Showcase > Cluster Events > Cluster Vehicles by Onboard Metrics

2. Explore the Cluster Number Events:

   What preprocessing steps is it using? What algorithm is it using?

3. Create a new experiment:

   Click Experiments > Create New Experiment

   Experiment Type: Cluster Numeric Events
   Experiment Title: Cluster Vehicles

   Click “Create”

4. Create a cluster using the dataset:

   Enter a search:
   ```
   | inputlookup mytrackdata.csv
   ```

   Preprocess method: StandardScaler
   Select the fields to preprocess: [Select all x_ fields]

   Click “Apply”

   Algorithm: K-means
   Fields to use for clustering: [Select all fields with “SS_x” prefix]
   K (# of centroids): 6 (the number of vehicles)

   Click “Cluster”

5. Save and publish the cluster experiment:

   Click “Save”

   Experiment Title = Cluster Vehicles

   Click “Go to Listings Page”

   Click “Publish” for Cluster Vehicles

   New Main Model Title: carcluster
   Destination App: Splunk Machine Learning Toolkit

   Copy the SPL snippet

Click “OK”

6. Use the SPL snippet in a new search:

```
| inputlookup mytrackdata.csv | apply [SPL snippet from step 5] | apply carcluster
```

7. Count values for the various vehicle types and average the cluster distance by cluster and save:

```
| inputlookup mytrackdata.csv | apply [SPL snippet from step 5] | apply carcluster
| stats count values(y_vehicleType) avg(cluster_distance) by cluster
```

Dashboard Title: ML Experiments
Panel Title: Cluster Statistics

8. Create some statistics:

```
| inputlookup mytrackdata.csv | apply [SPL snippet from step 5] | apply carcluster
| fit PCA k=3 SS_*
```

9. Rename some of the fields to translate the statistics into a 3D chart:

```
| inputlookup mytrackdata.csv | apply [SPL snippet from step 5] | apply carcluster
| fit PCA k=3 SS_* | rename cluster as clusterId, PC_1 as x, PC_2 as y, PC_3 as z
```

10. Create and save 3D Scatterplot to dashboard:

Select the “Visualizations” tab and ensure that “3D Scatterplot” is selected as the visualization type.

Dashboard Title: ML Experiments
Panel Title: 3D Cluster Map

Click “Save”