1 ML Tutorial 2

1.1 Predicting Categorical Fields

1. Select the "Splunk Machine Learning Toolkit App", and then select "Experiments". Next select "Predict Categorical Fields". Then create a new experiment and give it the name of "Firewall". Once created, then add the firewall traffic dataset:

```
| importlookup firewall_traffic.csv
```

How many records are there?

Next trim the dataset down to 50,000 records with:

```
| importlookup firewall_traffic.csv
| head 5000
```

Next we will use logistic regression to make a prediction. For the algorithm select "LogisticRegression", and then "used_by_malware" as the field to predict, and then select all the other fields for the fields to be used for predicting. Confirm that we are using 70% of the data to predict, and then select "Fit Model".

What are the values in the confusion matrix?
What is the Precision score:
What is the Accuracy score:
What is the F1 score:
What is the Recall score:

Next change the model to "SVM", and recompute your model:

What are the values in the confusion matrix?
What is the Precision score:
What is the Accuracy score:
What is the F1 score:
What is the Recall score:

Next change the model to "RandomForestClassifer", and recompute your model:

What are the values in the confusion matrix?
What is the Precision score:
What is the Accuracy score:
What is the F1 score:
What is the Recall score:

Try the other available models, and determine which one is the best for this dataset:

Which is the best model?

2. First select Prediction Categorical Fields, and give the experiment a name (such as cars). We first read the data in with:

```
| importlookup track_day.csv
```

Answer the following :

How many records are there?
What are the five cars defined:
What are the collected parameters:

Once populated, select logistic regression for your model. And then use the numeric fields to predict this, and where we will use 70% of the data to train, and then 30% of the data to test our prediction.

Precision:
Recall:
Accuracy:
F1:

And so while the success rate for true positives is fairly good, next use Random Forest Classifier (and which is made up from a number of models). What are the results:

Precision:
Recall:
Accuracy:
F1:

1.2 Predicting Numeric Fields

3. First select Prediction Numerical Fields, and give the experiment a name (such as cancer_experiment). We first read the data in with:

```
| inputlookup df.csv
```

Now train for the Cancer DR against the other parameters, and determine the best model to use:

Model 1:

R2 score:

Model 2:

R2 score:

Model 3:

R2 score:

Model 4:

R2 score: