Motivation

The Seattle Checkouts by Title consists of variables describing the book and the checkout. One variable describes whether the title is a physical copy or a digital copy. Since the dataset included time variables describing the checkout, I wondered whether the year of checkout, the month of checkout, and the number of times checked out were related to whether the title was a physical copy or a digital copy--maybe it is the case that digital copies are checked out at more recent dates due to increased use of technology in more recent times, or that physical copies were checked out a larger number of times because they seem to be more common. To see if the checkout year, checkout month, and number of times the title was checked out could predict whether the title was a digital copy or a physical copy, I performed Logistic Regression.

Code Snippet and Explanation: Getting the Data into a Usable Format

```
from pyspark.ml.feature import RFormula
rf = RFormula(formula="~ CheckoutYear +  CheckoutMonth + Checkouts")
final_df_rf = rf.fit(df).transform(df)
final_df_rf.show()
```

This code snippet vectorizes the regressors in the data by making them into a dense vector. I used the function RFormula, which mimics the equation format in logistic and logistic regression functions in R, to vectorize the regressors. This function allowed for vectorization to be more efficient, simpler, and coded in fewer lines than creating and applying a User Defined Function.

```
from pyspark.ml.feature import StringIndexer
indexer = StringIndexer(inputCol="UsageClass", outputCol="UsageClassBinary")
indexed = indexer.fit(final_df_rf).transform(final_df_rf)
indexed.show()
```

This code snippet converts the text data in the UsageClass column into binary 0 or 1 to represent the two categories to allow it to be used as a variable in Logistic Regression.

Visualization

With an area under the ROC curve of 0.500 and an area under the PR curve 0.503, the model consisting of the regressors seems to be a bad predictor of whether the item checked out is physical or digital.