Machine Learning Introduction

Agenda - Today

- Moneyball example background
- Regression
- Exploratory Data Analysis – 99 Wins

No Free Lunch Theorem

You don’t know which model will perform optimally for a given problem when you start.

Optimally can include

- Accuracy
- Speed
- Precision
- Recall
- Etc

EDA – Frame the Question

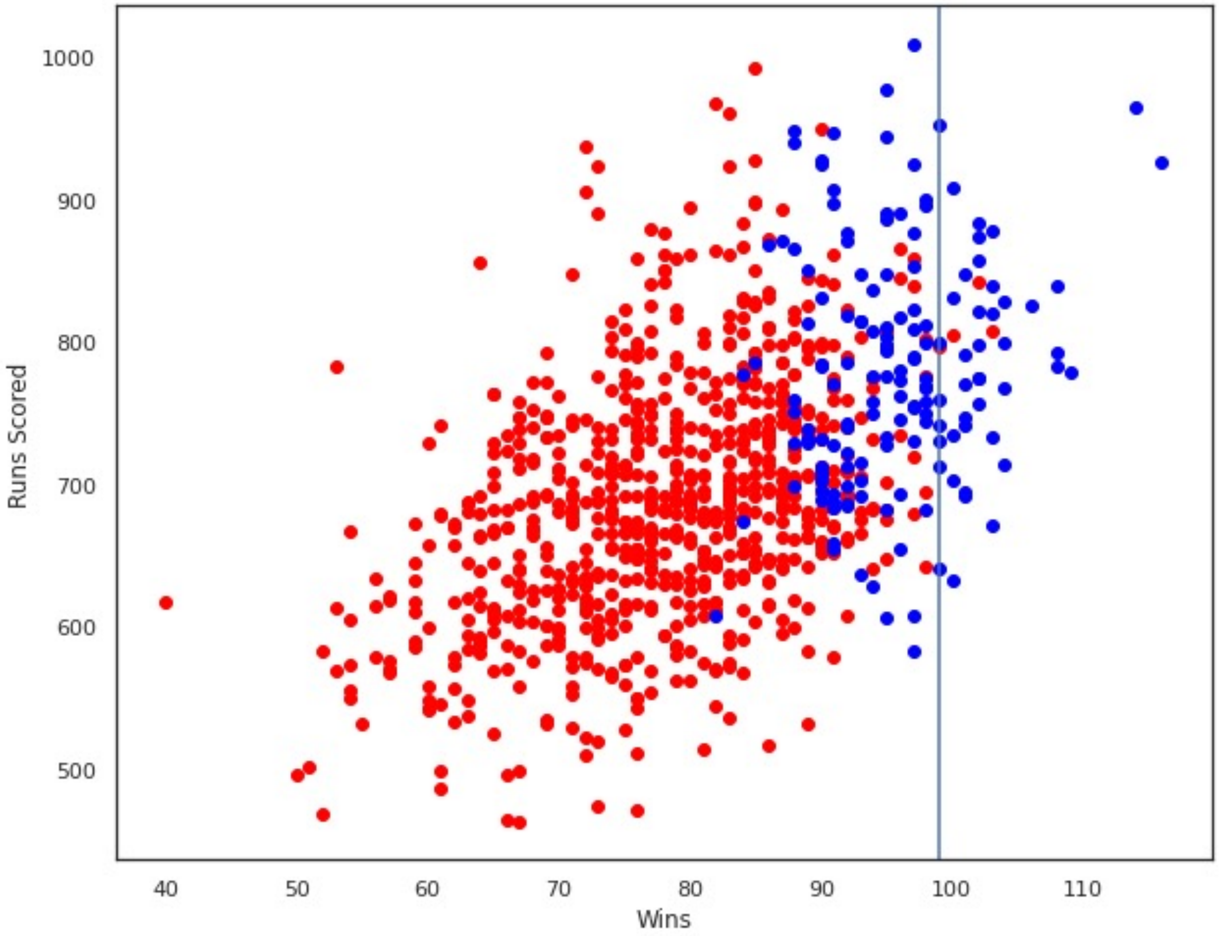
Moneyball

- Goal: make the playoff
- How?

EDA – Frame the Question

Moneyball – how?

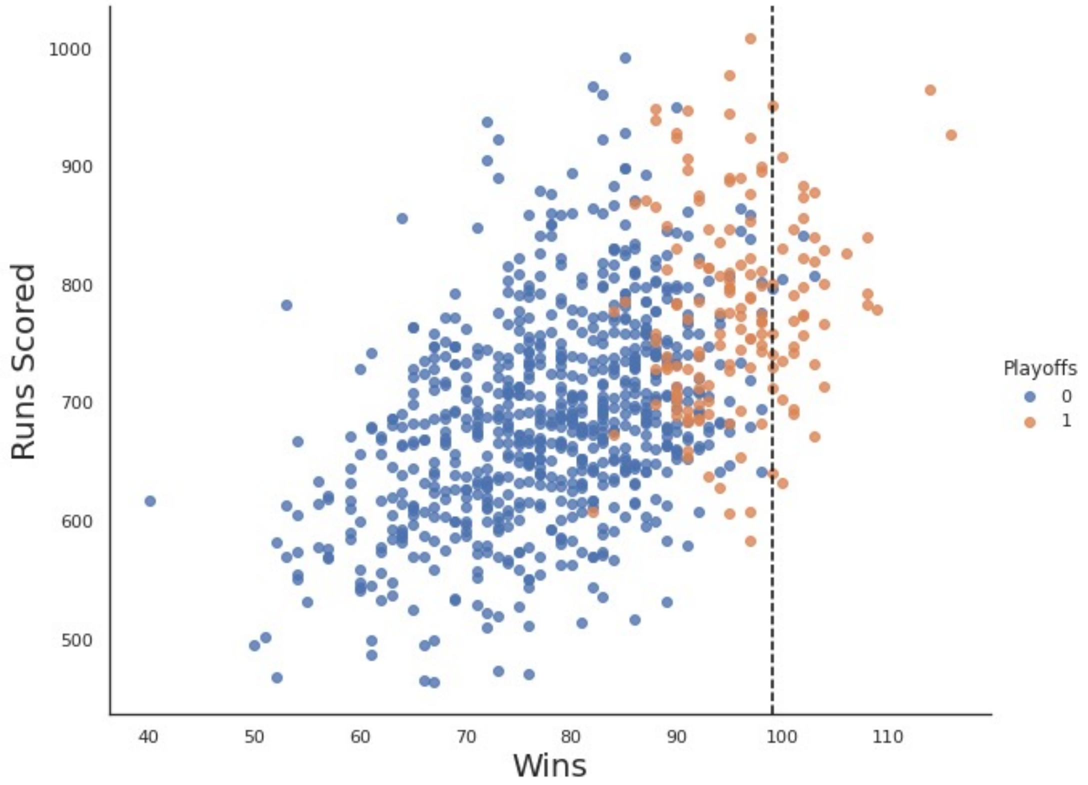
- Win 99 games
- How do you win 99 games?



EDA – Frame the Question

Moneyball

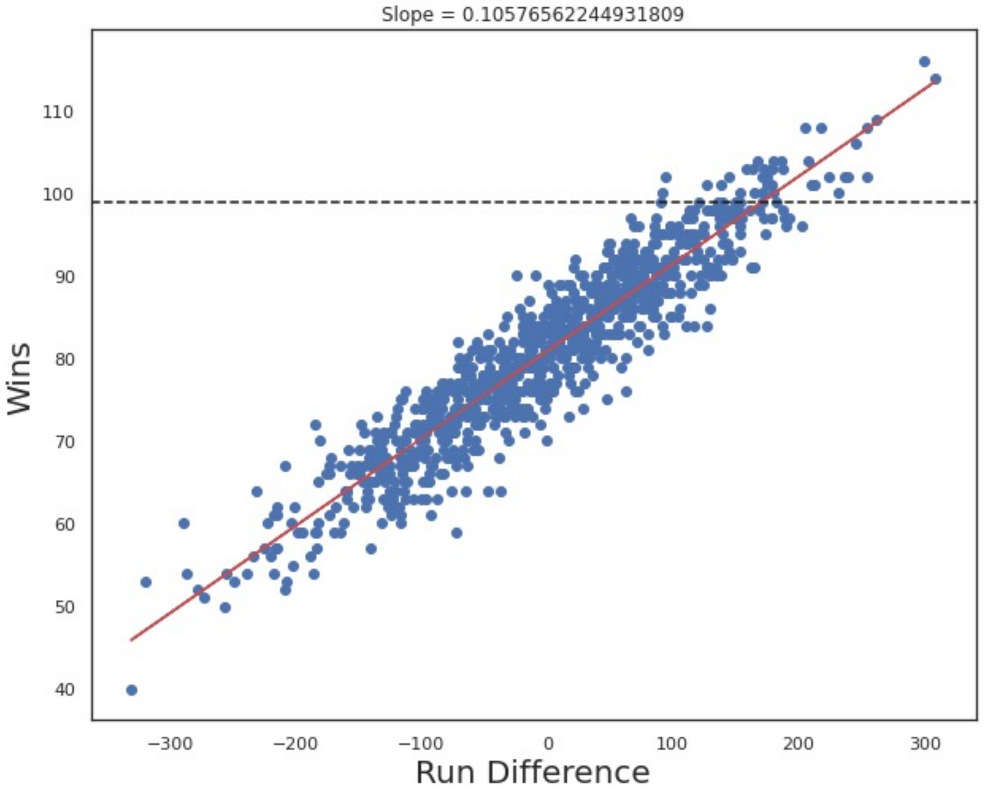
- How do you win 99 games?
- Score runs!
- How many?



EDA – Frame the Question

Moneyball

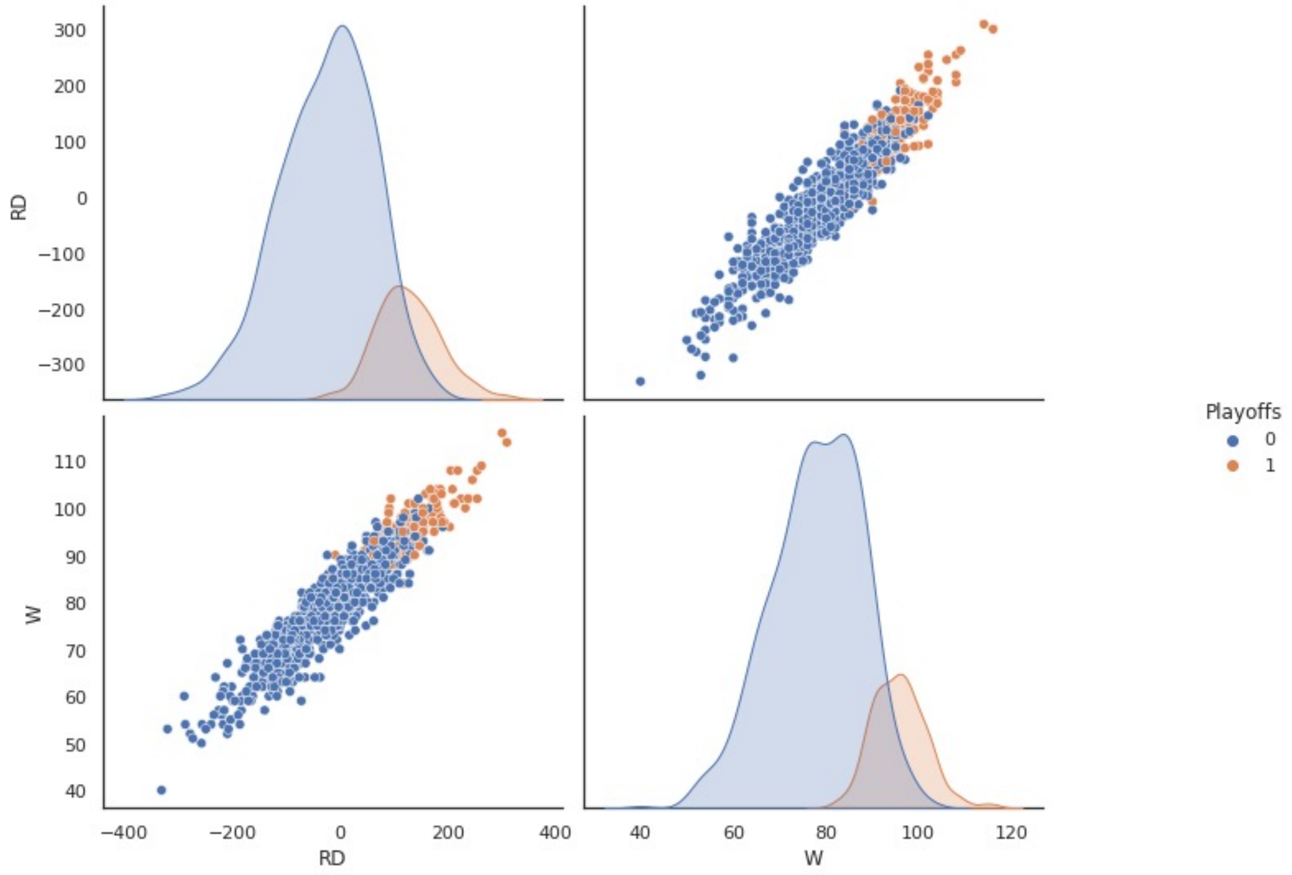
- How many?
- More than your opponent!
- Specifically, about 180 more



EDA – Frame the Question

Moneyball

- How many?
- More than your opponent!
- Specifically, about 180 more



Regression– Build a team

Look at code for story: MoneyBallStory

Next, look at general code for team building: detailedMoneyball

- Need more data
  - Position
  - Times at bat
  - Bat stats by type of results: single, double,base on balls, etc.
  - Salary

Regression– Measures

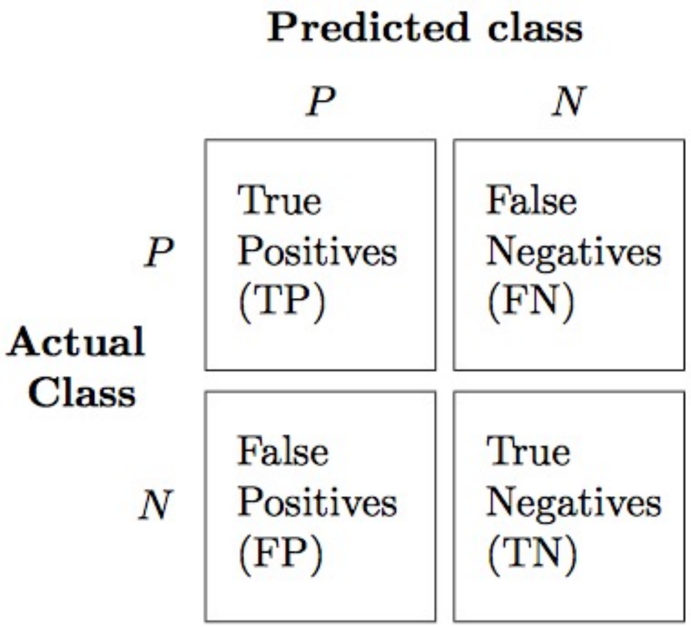
Most common performance measure for regression is R-Squared, the amount of the variability in the data that is explained by the model

Classifiers

While regression is associated with numbers, classifiers are associated with categories such as TRUE/FALSE, FRAUD/NOTFRAUD, GREEN/BLUE/RED.

Most common is binary classifier which can be built upon to make multiple category classifiers through repeated fits

Classifier Performance Analysis

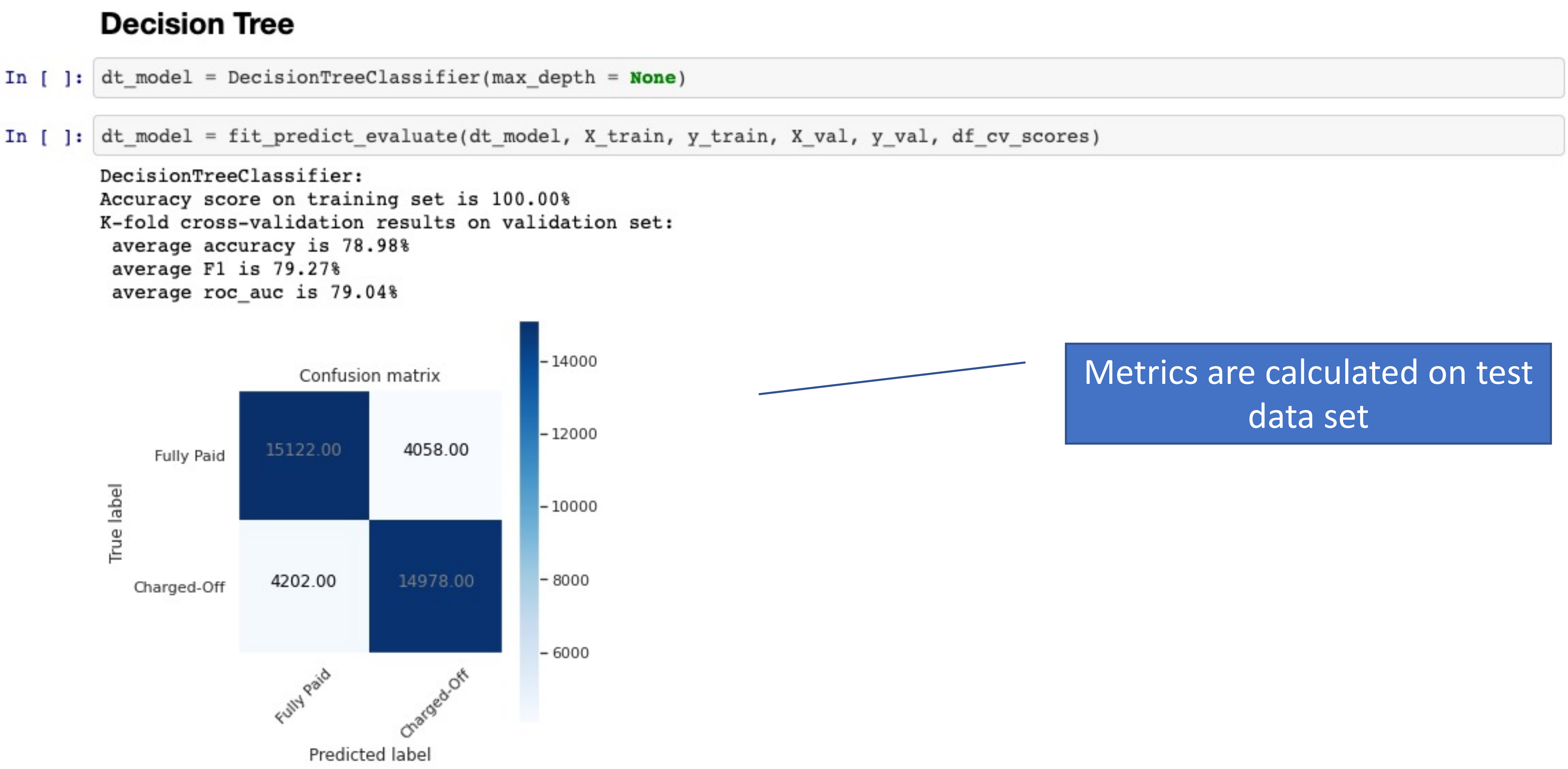


		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

True positives (TP), predicted positive and it was in fact positive!

True Negatives (TN), predicted negative and in fact it was negative!

Loans Modelled with Decision Tree



Decision Tree

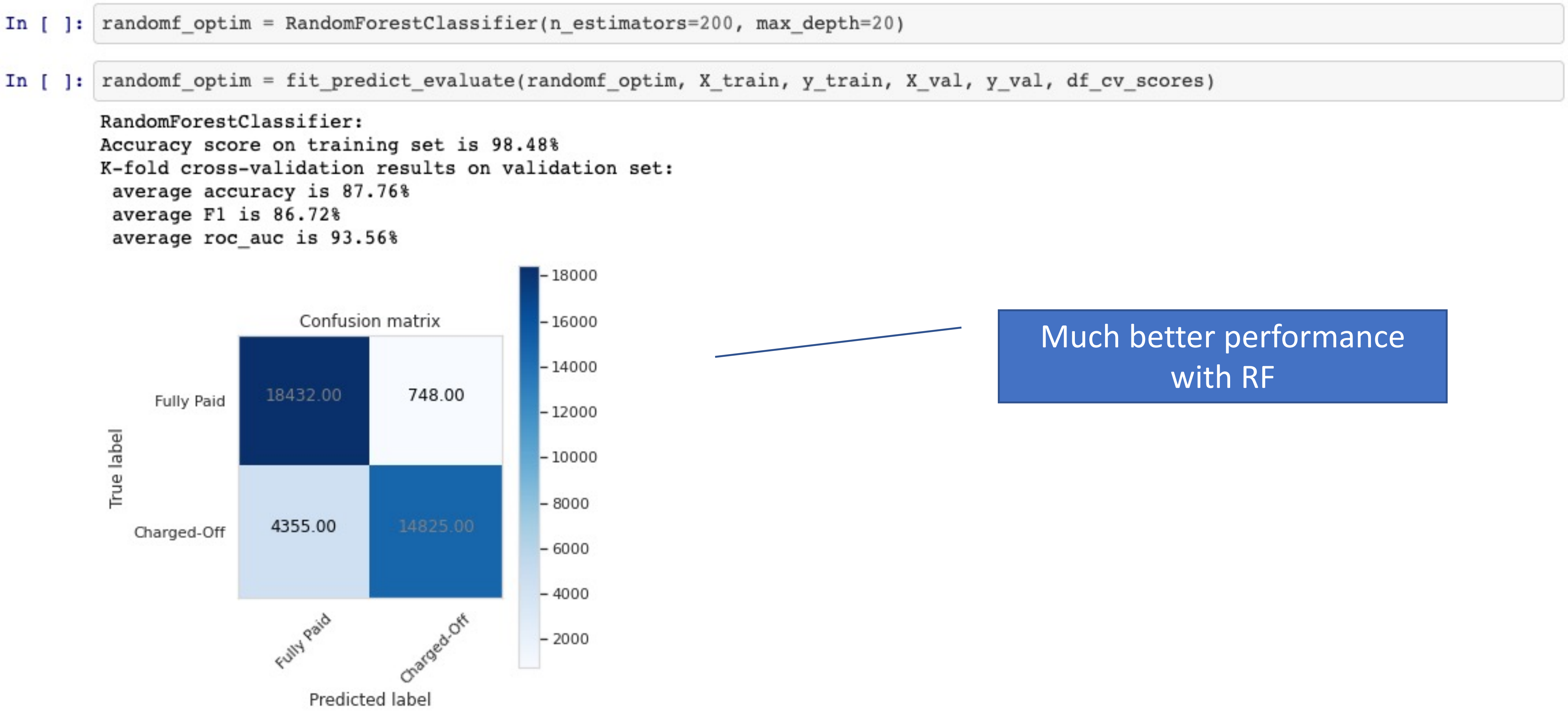
```
In [ ]: dt_model = DecisionTreeClassifier(max_depth = None)
```

```
In [ ]: dt_model = fit_predict_evaluate(dt_model, X_train, y_train, X_val, y_val, df_cv_scores)
```

```
DecisionTreeClassifier:
Accuracy score on training set is 100.00%
K-fold cross-validation results on validation set:
 average accuracy is 78.98%
 average F1 is 79.27%
 average roc_auc is 79.04%
```

Metrics are calculated on test data set

Loans Modelled with Random Forest



```
In [ ]: randomf_optim = RandomForestClassifier(n_estimators=200, max_depth=20)
```

```
In [ ]: randomf_optim = fit_predict_evaluate(randomf_optim, X_train, y_train, X_val, y_val, df_cv_scores)
```

```
RandomForestClassifier:
Accuracy score on training set is 98.48%
K-fold cross-validation results on validation set:
 average accuracy is 87.76%
 average F1 is 86.72%
 average roc_auc is 93.56%
```

Much better performance with RF

Summary

- EDA helps frame the objective
- Hold out data to test against
- Use multiple models to identify best approach