PREDICTING SEATTLE PUBLIC LIBRARY CHECKOUT RECORDS
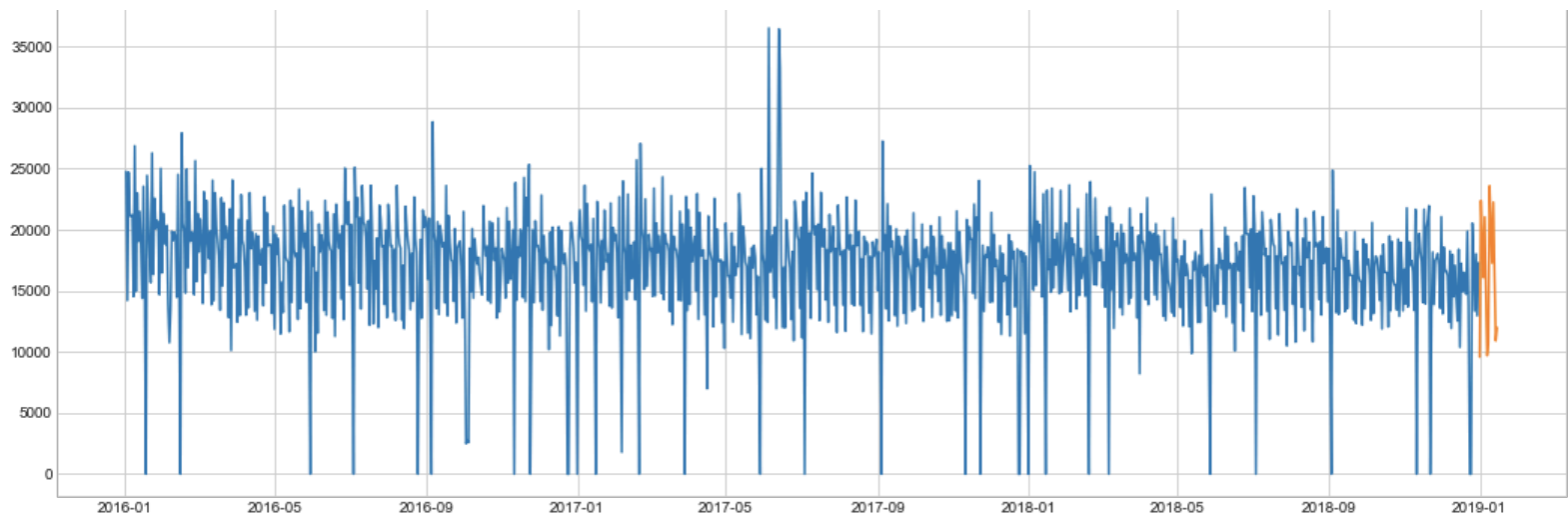
William Acosta

Problem Definition

How can we use publicly available library data to analyze borrowing trends and predict future checkouts?

Results could influence decisions:

- Staffing

- Procurement

- Branch focus

Dataset Description

Checkout records

- https://data.seattle.gov/dataset/Checkouts-by-Title-Physical-Items-/3h5r-qv5w

- About 19.2 million rows by 6 columns (~ 1.4 GB)

- BibNumber, Checkout datetime, call number, etc.

Library Catalogue

- https://data.seattle.gov/Community/Library-Collection-Inventory/6vkj-f5xf

- 2,687,149 rows by 13 columns (852 MB)

- BibNumber, bibliographical info, category info, item count, etc.

EDA findings

- In 2016, the Top 10 checkouts were all DVDs
- Technology books are by far the most popular adult non-fiction subject
- Saturday is the busiest day at the library
- Friday and Sunday are the slowest days
- 75% of checkouts are from materials released in past 8 years; 25% are from the last 2 years

# EDA findings

In 2016, the Top 10 checkouts were all DVDs

	Title	Format Subgroup
1	The Martian [videorecording] / Twentieth Centu...	Video Disc
2	Bridge of spies [videorecording] / Touchstone ...	Video Disc
3	The big short [videorecording] / Paramount Pic...	Video Disc
4	Spotlight [videorecording] / directed by Tom M...	Video Disc
5	Inside out [videorecording] / Disney ; Pixar A...	Video Disc
6	SPECTRE [videorecording] / Metro Goldwyn Mayer...	Video Disc
7	Brooklyn [videorecording] / Fox Searchlight Pi...	Video Disc
8	Trainwreck [videorecording] / Universal Pictur...	Video Disc
9	Jurassic world [videorecording] / Universal Pi...	Video Disc
10	Mission: Impossible. Rogue nation [videorecord...	Video Disc

EDA findings

Saturday is the busiest day at the library
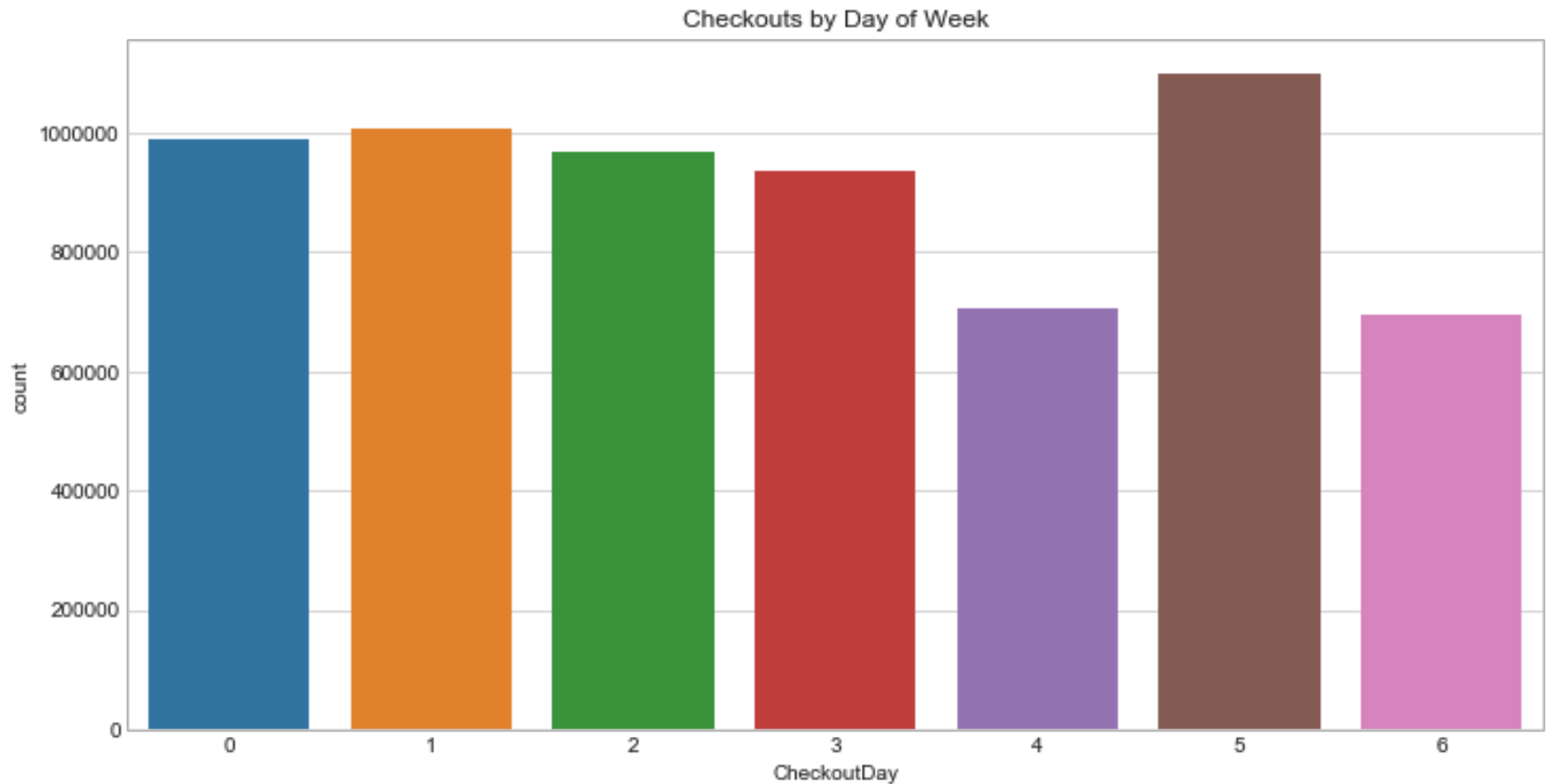
Trend decomposition
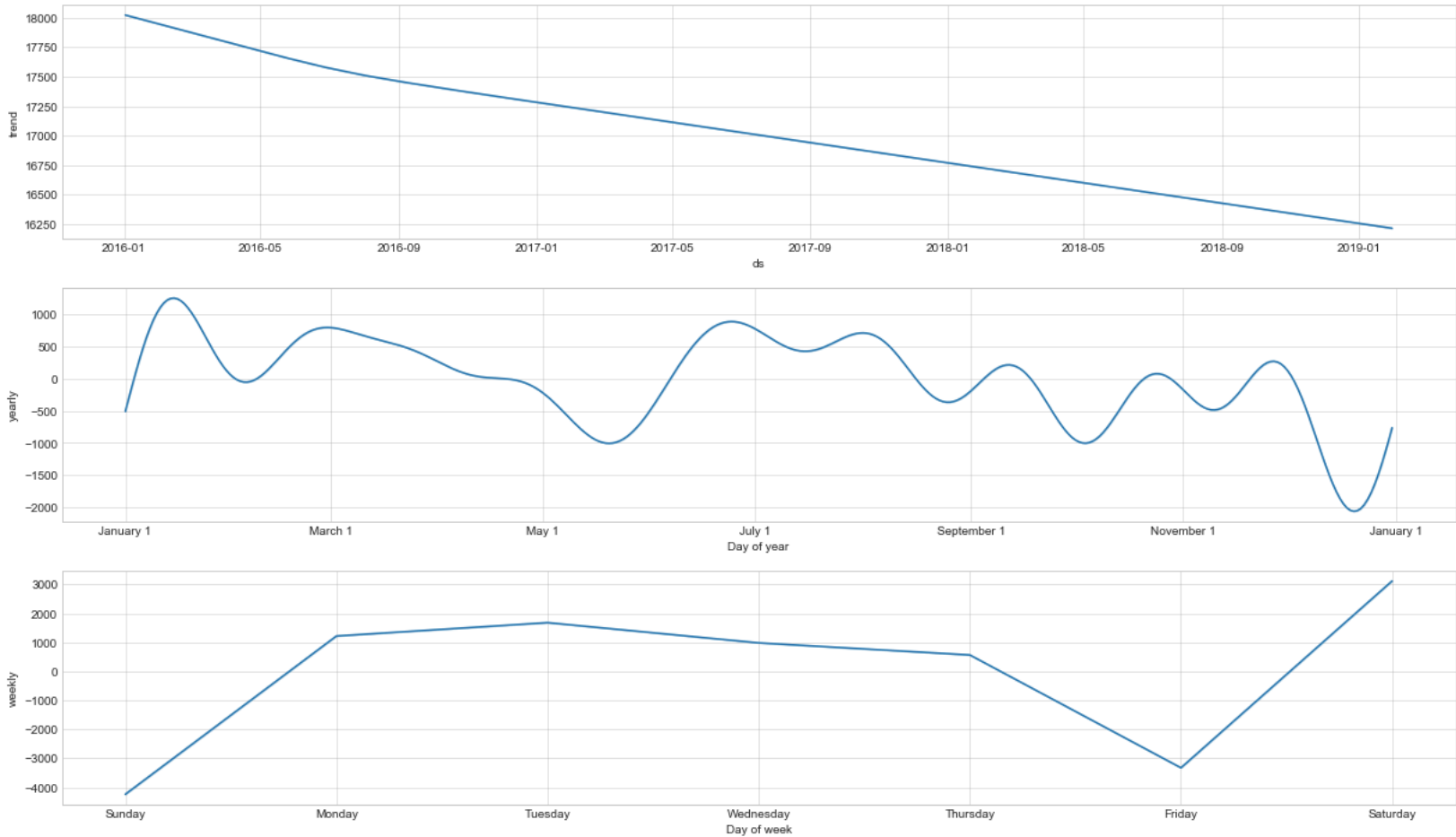
Approach

- Take 3 years of checkout data from 2016-2018 and transform it into an algorithm friendly tabular format

- Feature engineering
  - Create shifts for the checkout totals 7 days before a data point
  - Create calendar based features for months, day of week, holidays, days before and after holiday

- Use machine learning regression models to predict the next day total checkouts: Random Forest, KNN, Linear
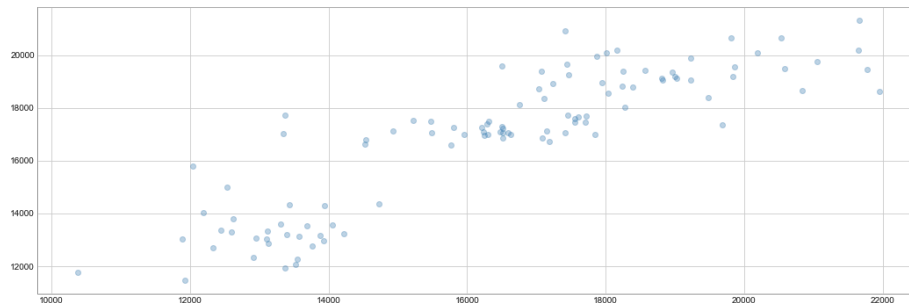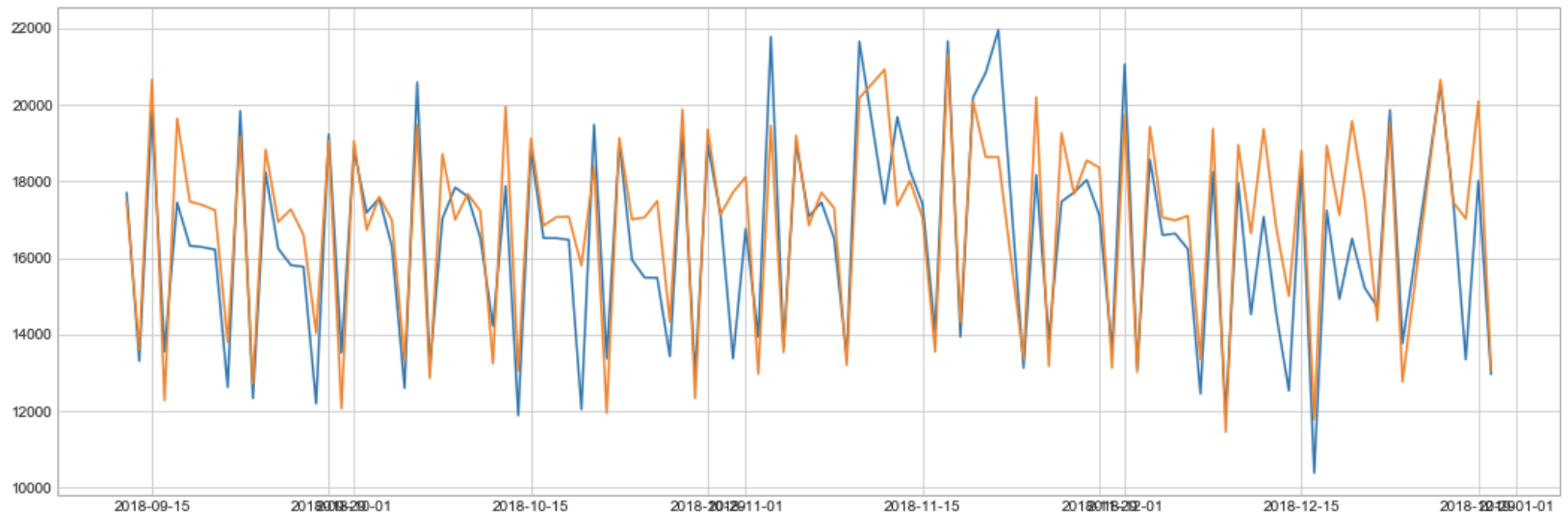
Results: Random Forest Regression
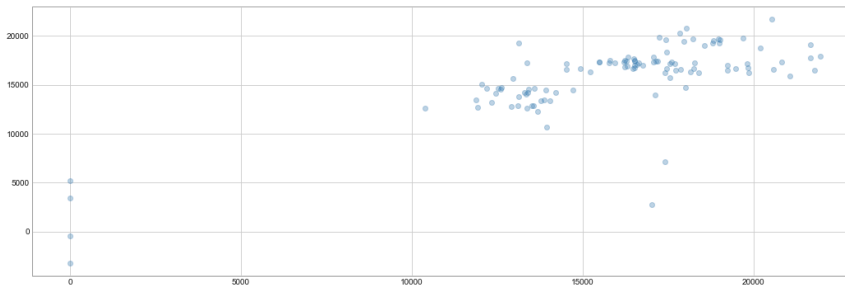
	model	mae	rmse
1	RandomForestRegressor50	1284.47	1899.16
0	LinearRegression	1753.27	2676.58
2	KNeighborsRegressor	1905.55	3969.34
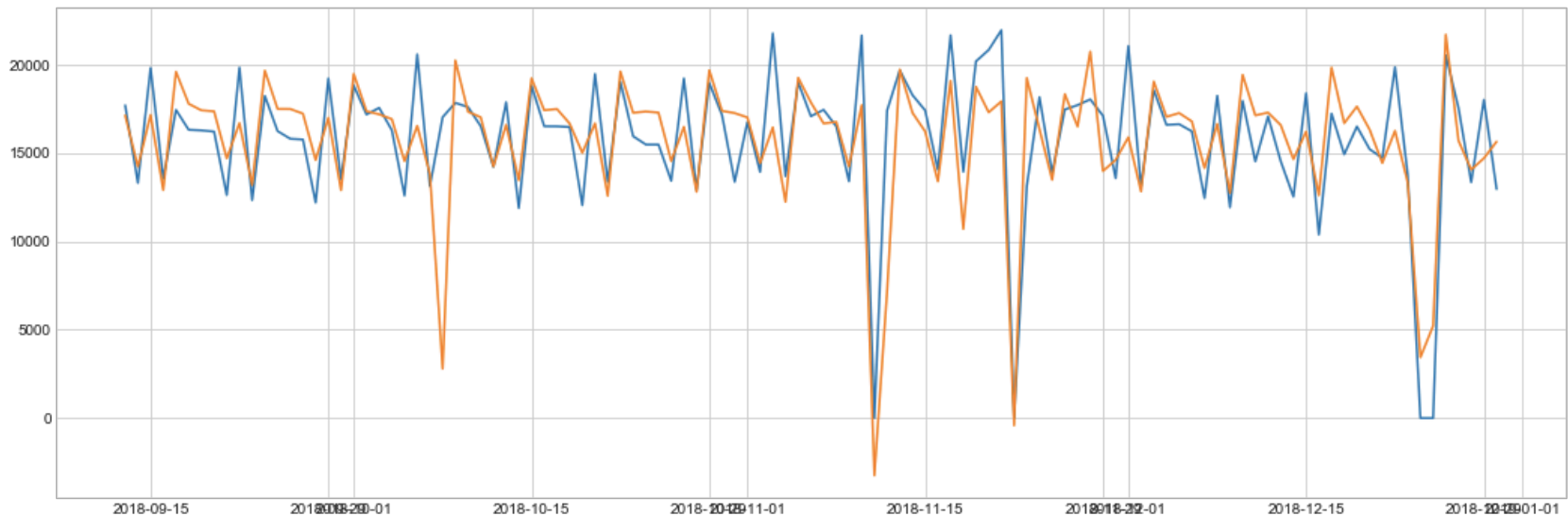
Results: Regression without outliers

	model	mae	rmse
1	RandomForestRegressor50	1064.31	1409.18
2	KNeighborsRegressor	1174.64	1611.77
0	LinearRegression	1608.94	2034.97

Prediction app

https://spl-checkout-predictor.herokuapp.com/