Library Book Use Prediction

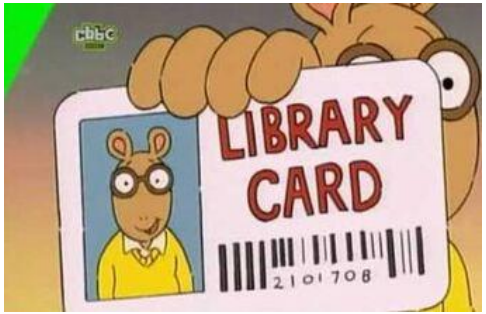
Item Checkout Prediction for Seattle Public Library

Group 19 - Andrew, Ash, Esme, Lori, Shishir

Agenda

- Background and Data
- Problem Description
- Related Work
- Pre-processing
- Machine Learning Outcome
- Run Time Performance
- Conclusion

Background: Seattle Library Data



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

Library Collection Management
How does it work???

Problem statement	Heuristic Approach	Improvements
Optimizing book selection / inventory management at libraries	Intuition of the librarian/static rules	Using past circulation data to predict future book checkouts

Goal

- For each item in the library's inventory, we want to predict whether or not the item will be checked out in the coming month

- Identify items less likely to be checked out and store off-site, or inform future book purchases

- 133,000 checkouts every month for an inventory of 760,000 unique items

Related Work

1996

ID3 Decision Tree[1]

Predicted checkouts using trees that can have up to 3 nodes at each split, trained using information gain criteria

2006

Impact of online reviews on book sales[2]

Study showed that online reviews have a statistically significant impact on book sales

2018

Identifying other important features for book sales[3]

ML models used to identify features that are most predictive of book sales - author's prior popularity, time of year etc.

Pre-processing

Data Pipeline

Title, ISBN, Floating Item, Item Barcode

- Lag based features (ex. number of checkouts last month)
- Number of books per Author
- Number of books per Publisher

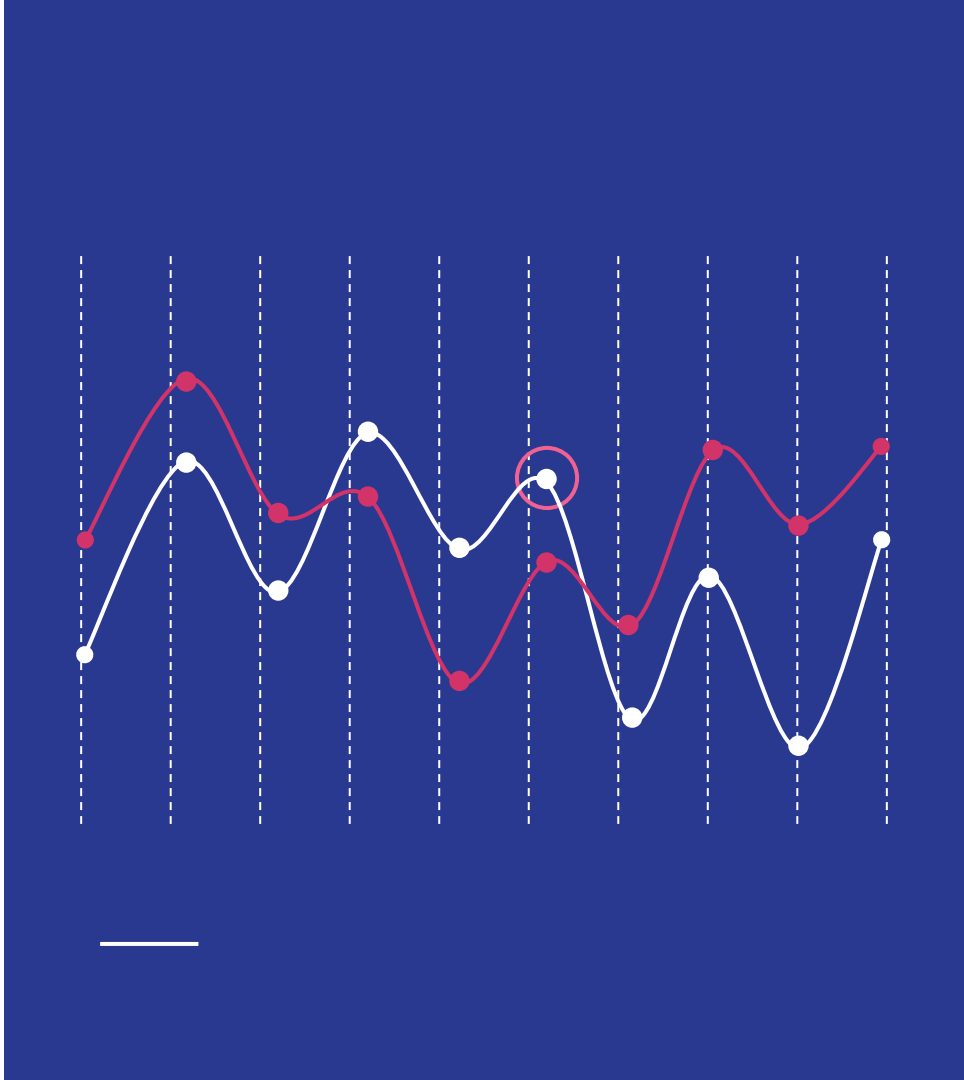
- Fill publishing year null values with mean value

- Cache
- MapPartitions
- Able to improve runtime by 5x

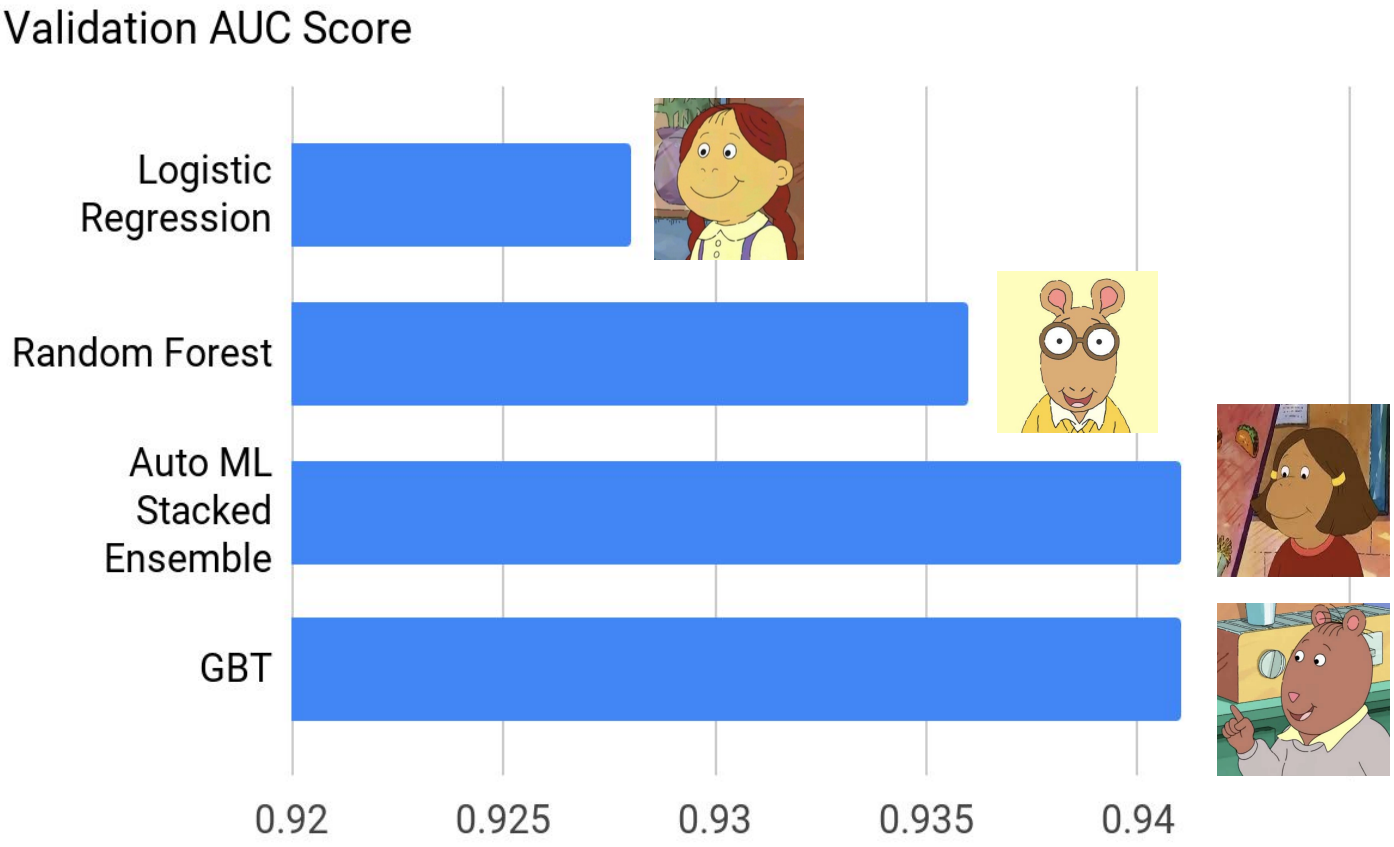
Out-of-time validation

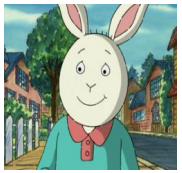
- Basic idea: Out-of-sample validation on a later dataset than the train data
- Train Dataset: All variables for data until November 2019, label = checked out or not (Yes/No) in November
- Test Dataset: All variables for data until December 2019, label = checked out or not (Yes/No) in December

Machine Learning Outcomes



Model Comparison

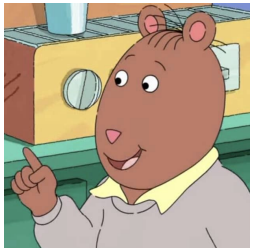




Multi layer perceptron: 0.507

😓

Cluster Comparison

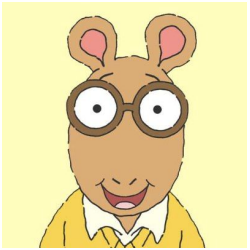


105s

m3.xlarge
2 instances

89s

m4.2xlarge
2 instances

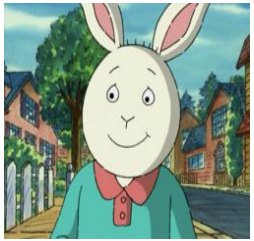


89s

m5.xlarge
2 instances

64s

m3.xlarge
3 instances



75s

m4.xlarge
3 instances

Takeaways

- We can accurately predict whether or not a book will be checked out or not in the next month
- Working on a smaller version of your dataset can be useful
- Feature engineering is valuable
- AWS works in mysterious ways 

Future Work

- Book category information - ex. “Young adult”
- Book type - fiction, non-fiction etc.
- Online reviews - ex. Goodreads
- Sales data
- Book renewal count
- Random hyperparameter search
- Deep Learning models

Thank you!


friend: how long did you sleep for?
me: 8
friend: hours?
me: minutes

Appendix: References

1. Craig Silverstein and Stuart M. Shieber. Predicting individual book use for off-site storage using decision trees. Library Quarterly, 66(3):266-293, July 1996. University of Chicago Press
2. Mayzlin, D. & Chevalier, J. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. SAGE Journals. doi: https://journals.sagepub.com/doi/10.1509/jmkr.43.3.345.
3. Yucesoy, B., Wang, X., Huang, J. et al. (2018) Success in books: a big data approach to bestsellers. EPJ Data Sci. 7, 7 doi:10.1140/epjds/s13688-018-0135-y