Automating the Moving Estimate Process

Ben Shaver, Data Scientist


KEILLOR
PONTOON

Location 1 (Starting)

Address	
City	
State	DC
Zip Code	
Type of Place	Please Select...
Do you know the square footage, ballpark?	
Length of walk from the truck to your front door? (10 ft ≈ 1 car length)	Please select...
How many stairs?	Please select...
Elevator?	Please select...
Are there any time restrictions we should know about?	
What type of parking is available for us?	○ Street ○ Loading Dock ○ Parking Lot ○ Driveway

The Moving Estimate

Moving Estimate

Time and Date: 8am on Friday, November 10th
Specs: 5 movers and a 26' truck (includes moving pads, handtrucks, dollies, and other necessary equipment)
Hourly Rate: $260.00/hour with a 3 hour minimum (any time beyond 3 hours is prorated in 15 minute increments)
Travel Fee: None
Time estimate: 3.5-4.5 hours
Total estimated costs: $910-$1170, *minimum cost:* $780

Target variables include:

- Estimated move duration
- Travel fee
- Rate
- Number of movers
- Size of truck

Goals:

- Predict how long a move will take
- Replicate a human-produced estimate
- Understand when moves take longer than expected

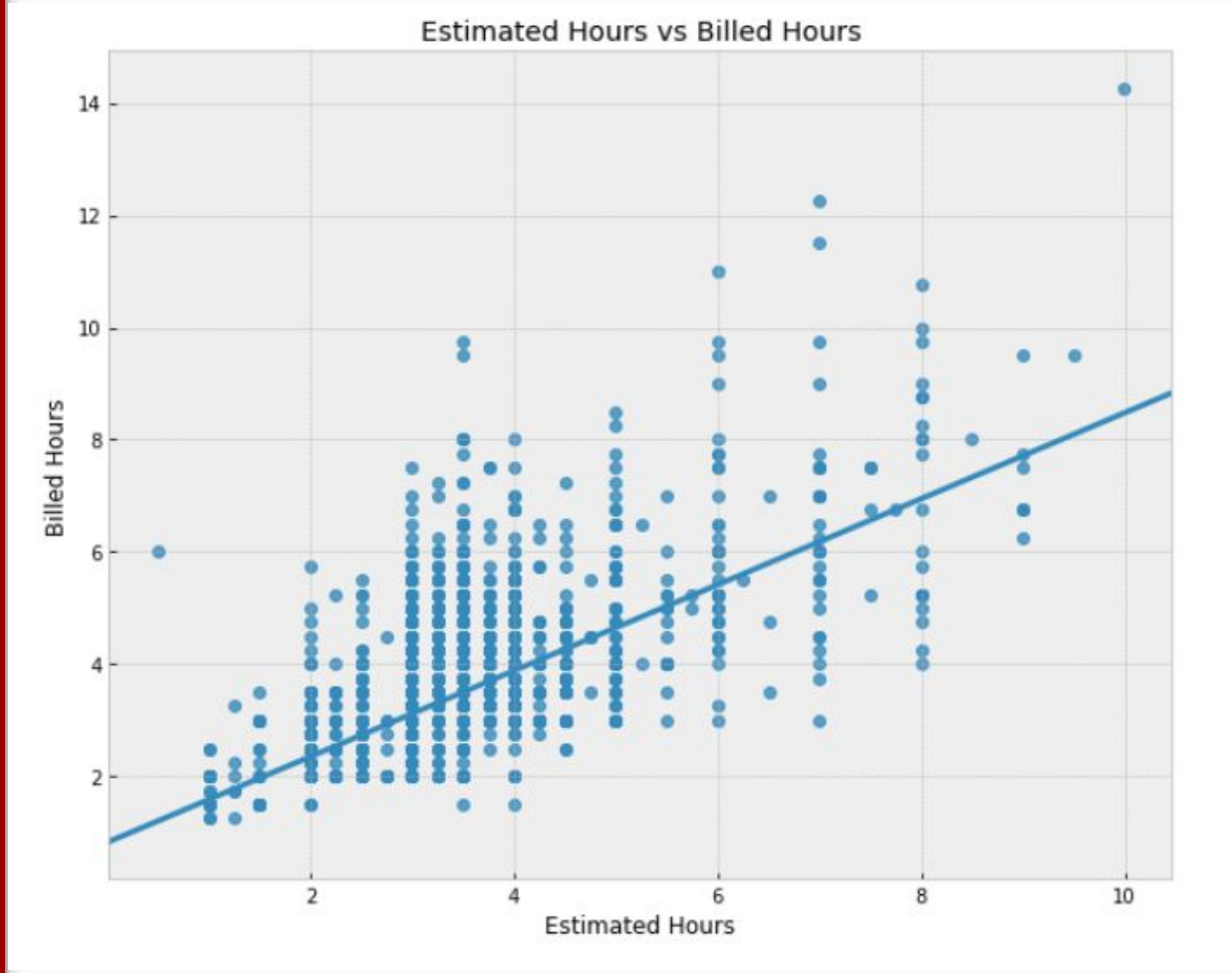
Hierarchical Model:

- First, classify a job as a ‘big truck job’ or a ‘small truck job’

VS

Hierarchical Model:

- Then, predict how long each move will take and what the estimate will be.



Criteria for Success:

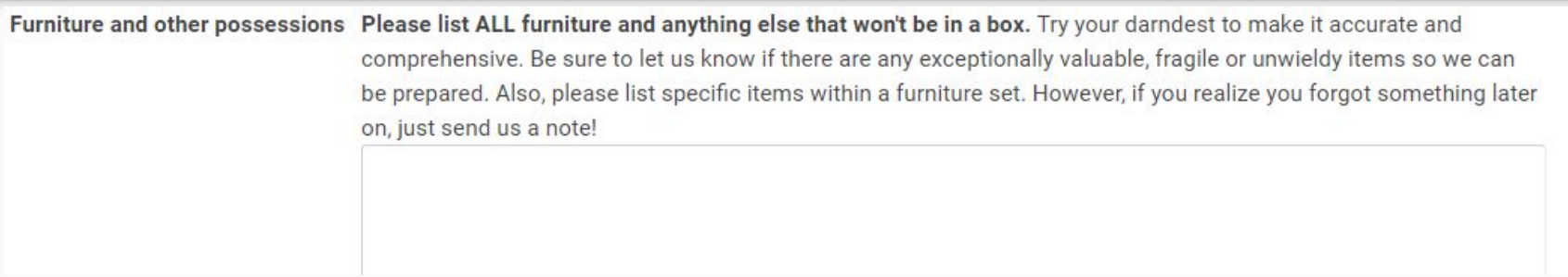
- Since we sent out big trucks all the time to complete jobs that could be performed by a smaller truck, my model should be highly sensitive to classifying ‘small truck jobs.’ If small truck jobs are are our true positive, we want a very high true positive rate.
- The BSM admin team is a ‘human estimator.’ Therefore, we can compute their R-squared. My model should aim to outperform the human estimator.

But first: data cleaning!

Types of Data:

- Target Variables:
  - Estimated hours, billed hours, size of truck.
- Numerical Variables:
  - Number of boxes and driving distance
  - At each location: square footage, stairs, number of rooms
- Categorical Variables:
  - Interest in packing services
  - At each location: parking type, building type, elevator type (if any)

Text Variables



- *Furniture*: listing
- *Reference*: how they heard about us
- *Mention*: anything else the client wants to mention

NLP Part One

- Sentiment Analysis:
  - Polarity: How positive or negative a text is
  - Subjectivity: How ‘fact-based’ versus ‘opinion-based’ a text is
  - Compute for each text field
- Part-of-Speech Tagging:
  - Counting the number of noun phrases in the furniture listing
- Hard Coding of NLP Features
  - IE, counting the raw length (in characters) of the furniture listing
  - Does the estimate contain a piano?
  - Searching for digit and noun phrase pairs, such as ‘3 bookshelves’
  - (The possibilities are endless)

NLP Part Two

- Tokenizing the furniture listing:
  - Selecting only words, not numbers
  - Dropping stop words
  - Lemmatizing
    - Turns ‘bookshelves’ into ‘bookshelf,’ which is problematic
- Constructing count vectors
  - Counting the occurence of the top 100 words:
    - *Table, chair, small, bed, dresser, tv, large, desk, queen, dining...*

Geocoding

- BSM has recently started using the Google Maps API to estimate driving time
- Available geographic data:
  - Geocoded addresses.
  - Latitude and longitude
  - Client-provided addresses (the last resort)
- Google Maps Distance Matrix API provides driving time between locations
- But:
  - People often move from ‘Alexandria, MD’ to ‘Washington, VA.’
  - Or they live ‘near Dupont circle’ in ‘NE’ DC.

Models

*I have *binders* full of mode

Models

- Small truck or big truck classification
- Small truck regression
- Big truck regression
- ‘Over or under’ classification

Model Evaluation

- The small/big model achieves 90+% true positive rate in classifying small truck jobs, based on a validation set.
- The small truck model achieves an R-squared in the same ballpark as the admin team, or 'human estimator.' About 28% vs 30%.
    - Consider a better metric?
- The over/under model has AUC-ROC of .7. Not bad considering it is predicting what humans miss.

Insights

- Most important features in over/under model:
  - Driving distance
  - Square footage
  - # characters in furniture list
  - Polarity and subjectivity of furniture list
  - # boxes, # noun phrases
  - Sentiment of *mention* text
- These are all variables the human estimators can pay closer attention to!

Insights

- Most important features in small/big model:
  - Square footage (at each location)
  - # boxes
  - # characters in furniture list
  - # noun phrases

Future Work

- Lemmatization turns ‘bookshelves’ into ‘bookshelf’ and can’t tell the difference between ‘5 couches’ and ‘a couch’
- Hard code to count number of common furniture types
    - Really replacing humans here
- Consult with stakeholders and current admin team to identify new predictors
    - E.g. clients who are black- or grey-listed, or who are Yelpers
- DATA CLEANING

Questions?

benpshaver@gmail.com

https://github.com/bpshaver/BSM_Analytics/

https://www.bookstoremovers.com/