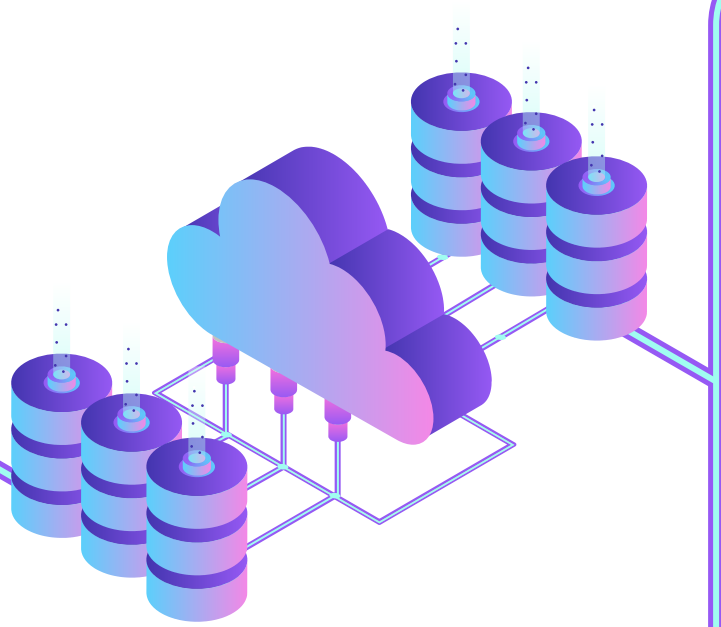
Predicting Ratings Based on Amazon Reviews

Logan King, Lori Chiu, Sarah Torrence

Problem Statement

Problem

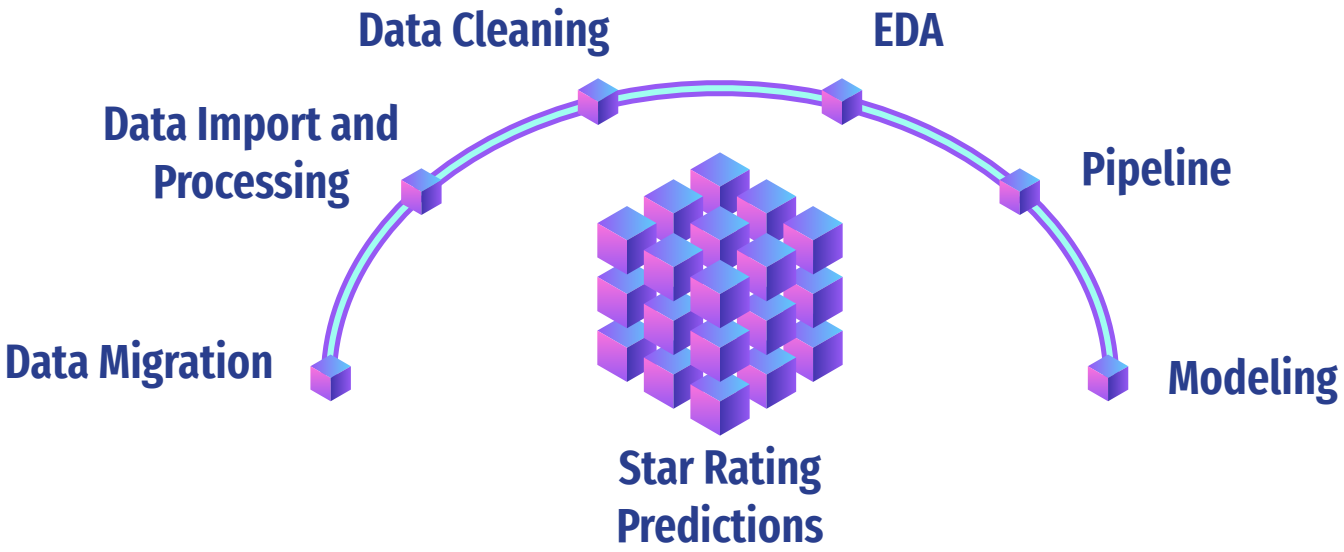
Amazon has an extensive database of reviews for each of their products. Consumers are using these reviews to decide whether or not to make a purchase. We want to understand the helpfulness of these reviews by predicting star ratings based on the text of the reviews.



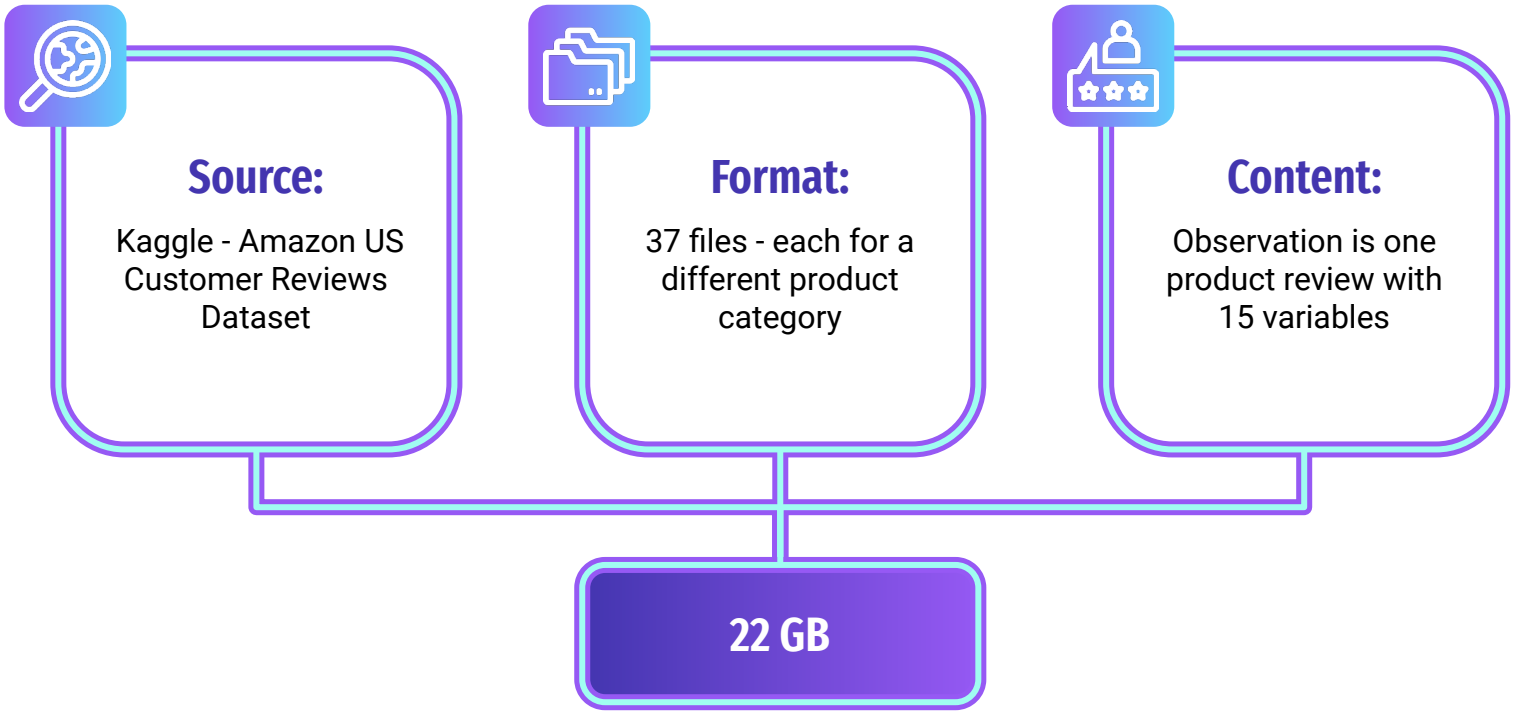
Application

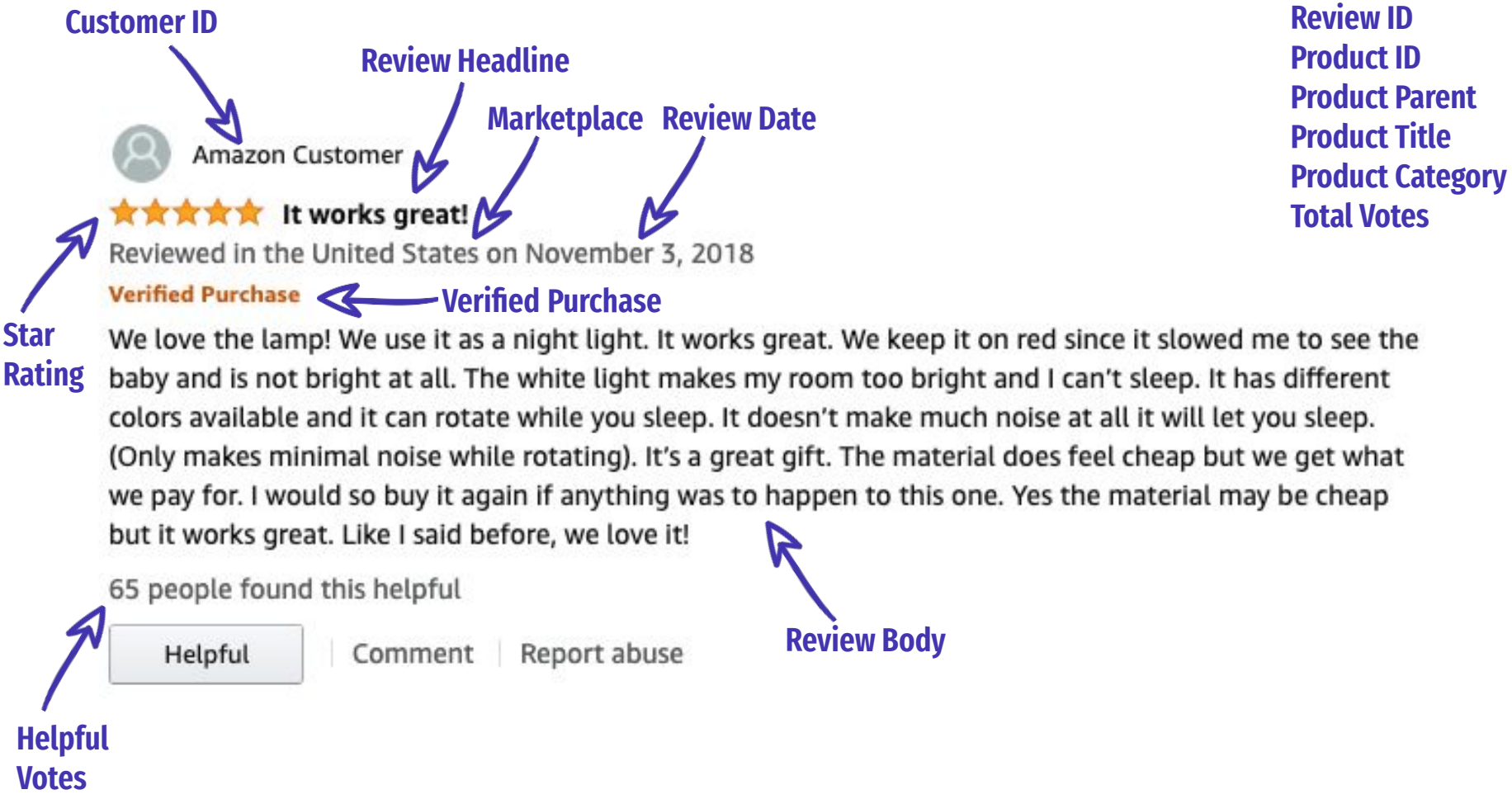
Amazon can use this model to predict star ratings for reviews that did not have any stars associated with the review. Amazon can accurately assign a star rating based on the text rather than the sentiment of the user.

Methodology



Data


Source:
Kaggle - Amazon US Customer Reviews Dataset
Format:
37 files - each for a different product category
Content:
Observation is one product review with 15 variables
22 GB


Customer ID
Review Headline
Marketplace
Review Date
Review ID
Product ID
Product Parent
Product Title
Product Category
Total Votes
Amazon Customer
It works great!
Reviewed in the United States on November 3, 2018
Verified Purchase
Verified Purchase
Star Rating
We love the lamp! We use it as a night light. It works great. We keep it on red since it slowed me to see the baby and is not bright at all. The white light makes my room too bright and I can't sleep. It has different colors available and it can rotate while you sleep. It doesn't make much noise at all it will let you sleep. (Only makes minimal noise while rotating). It's a great gift. The material does feel cheap but we get what we pay for. I would so buy it again if anything was to happen to this one. Yes the material may be cheap but it works great. Like I said before, we love it!
65 people found this helpful
Helpful
Comment
Report abuse
Review Body
Helpful Votes

Data Migration

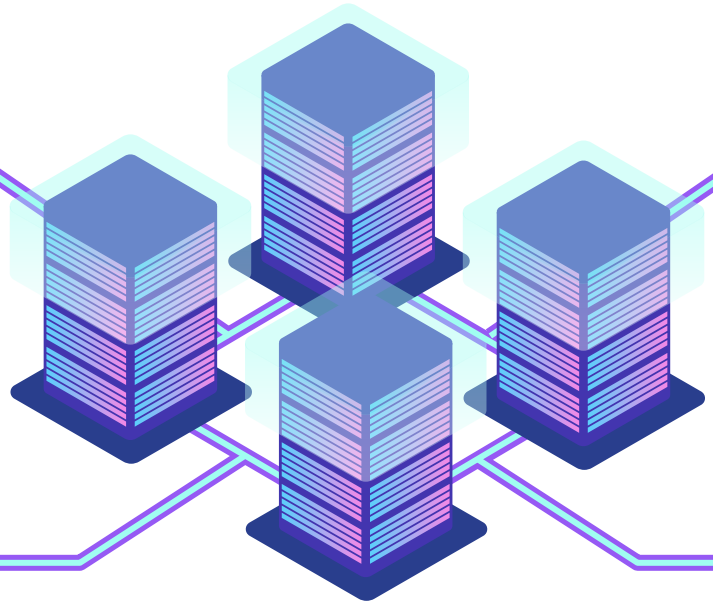
Data Import and Processing

Multiprocessing

Parallelized across 8 CPU

Pyspark

Converted to Pyspark DataFrame



Pandas

Imported as DF - filtered & cleaned

ACCRE

Interactive Jupyter Notebook

Data Cleaning

Remove NAs
Removed rows that contained NAs in review title or body

Remove Duplicates
Removed duplicate reviews

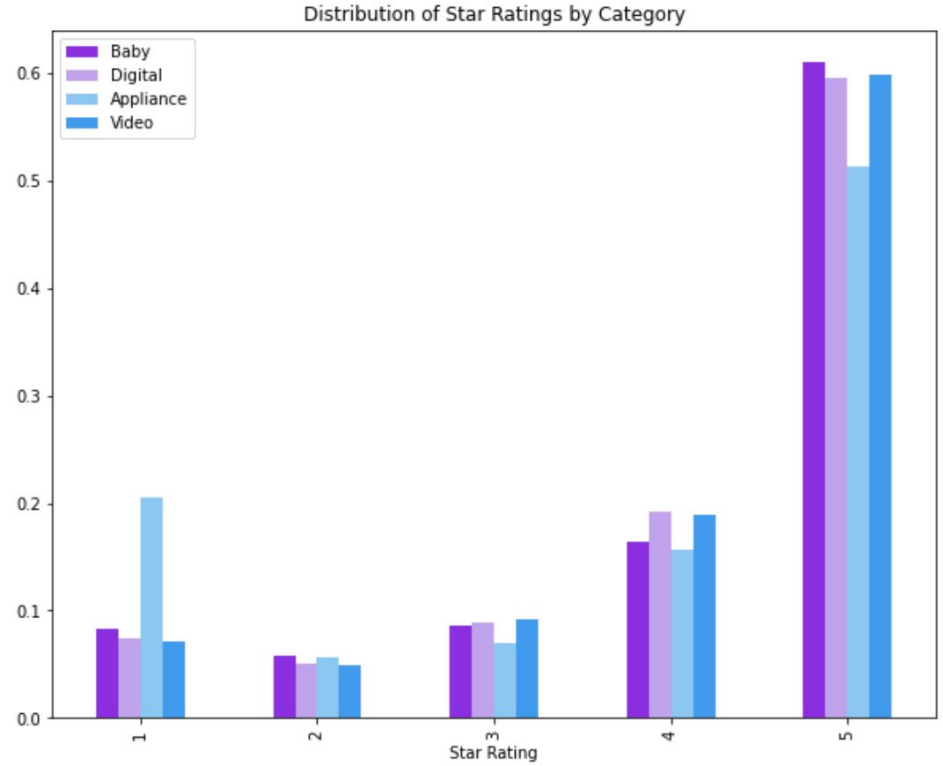
One-Hot Encode
One-hot encoded verified purchase variable

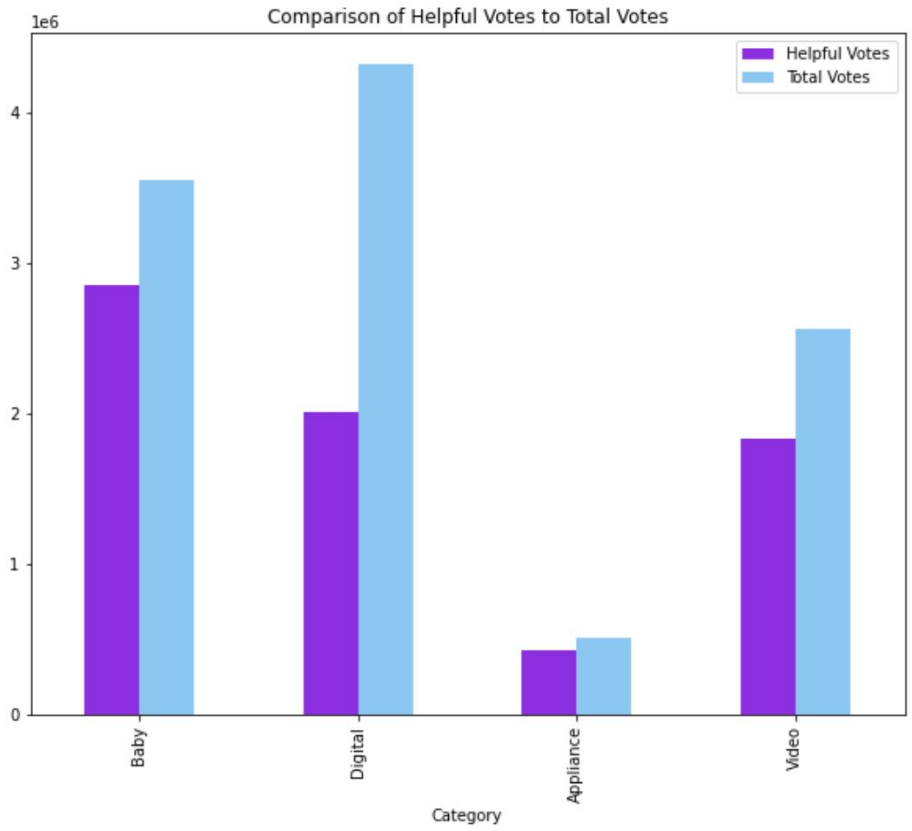
Drop Marketplace
All reviews are for US market

IDs
Unique identifiers for customers, reviews, and product

Product Category
Inconsistent values

EDA


Distribution of Star Ratings by Category
Baby
Digital
Appliance
Video
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
Star Rating


Comparison of Helpful Votes to Total Votes
1e6
Helpful Votes
Total Votes
0
1
2
3
4
Baby
Digital
Appliance
Video
Category

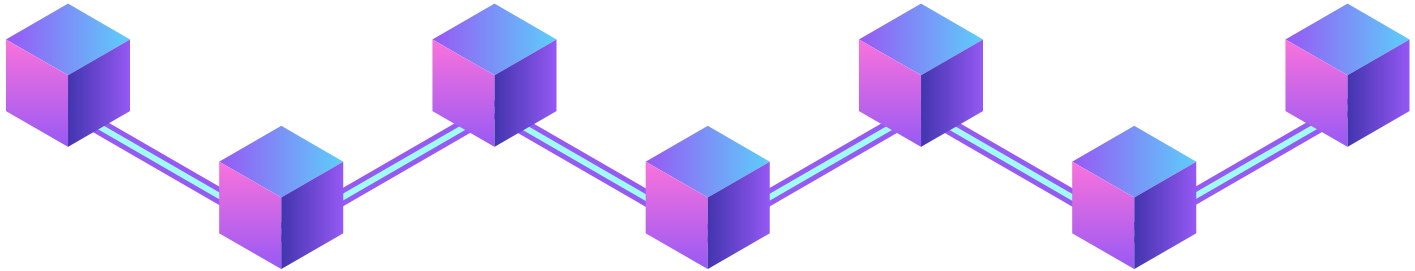
Data Features

Product Title

Product Category

Helpful Votes

Total Votes



Verified Purchase

Review Headline

Review Body

Star Rating

Pipeline

Transformation	Description	Variables
Tokenizer	Splits string inputs into a list of words	String
StopWordsRemover	Removes words classified as “Stop Words” from list input	String
Word2Vec	Transforms list of words into numeric vector	String
VectorAssembler	Collects all specified features into a single vector	All
Normalizer	Normalizes input set of numeric variables	Numeric

Modeling

Inputs

Star Rating, Normalized Features

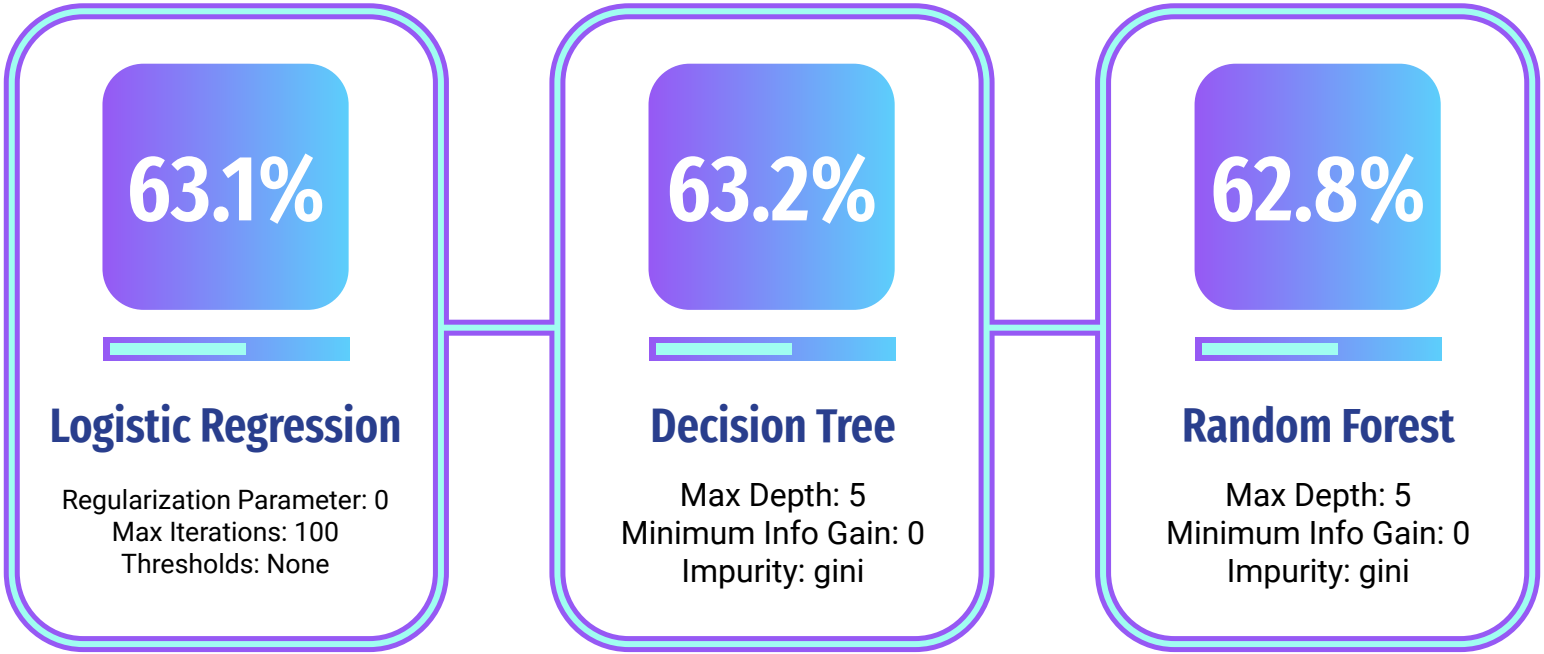
- Star Rating is the label input
- Normalized Features output by the pipeline are the features
- Train/Test Split: 70/30

Predictive Models

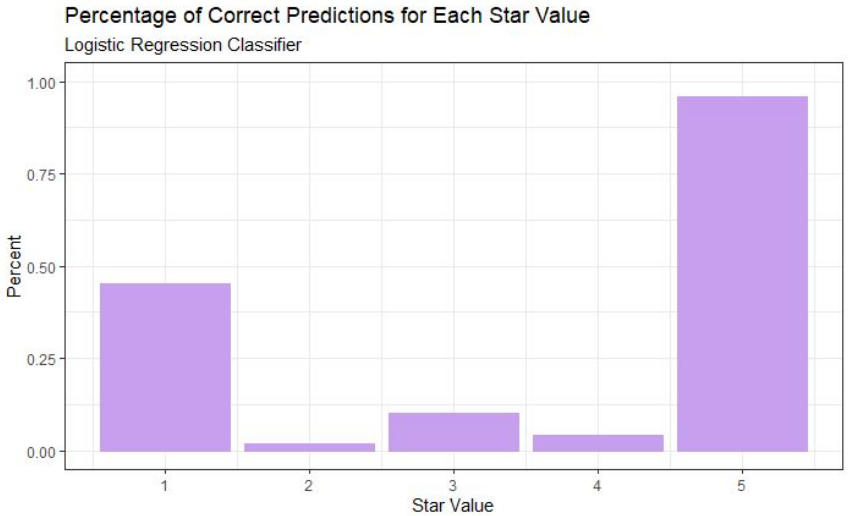
LR, DT, RF

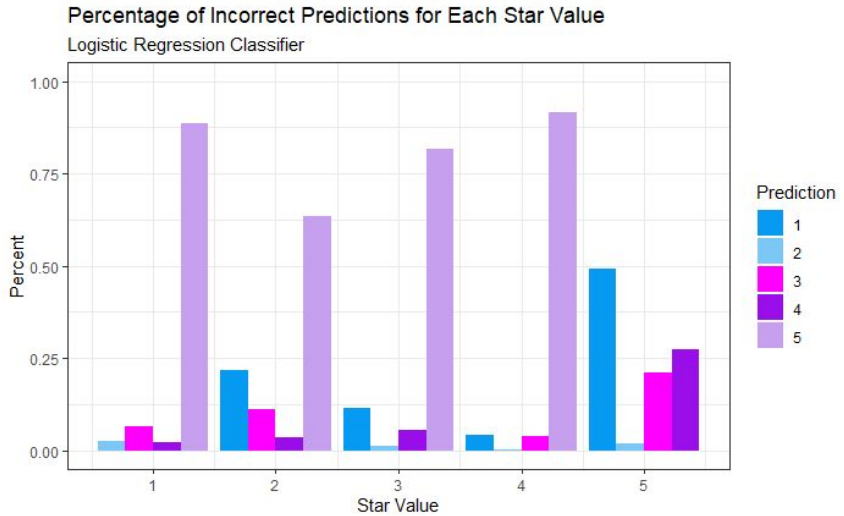
- Logistic Regression Classifier
- Decision Tree Classifier
- Random Forest Classifier

Results

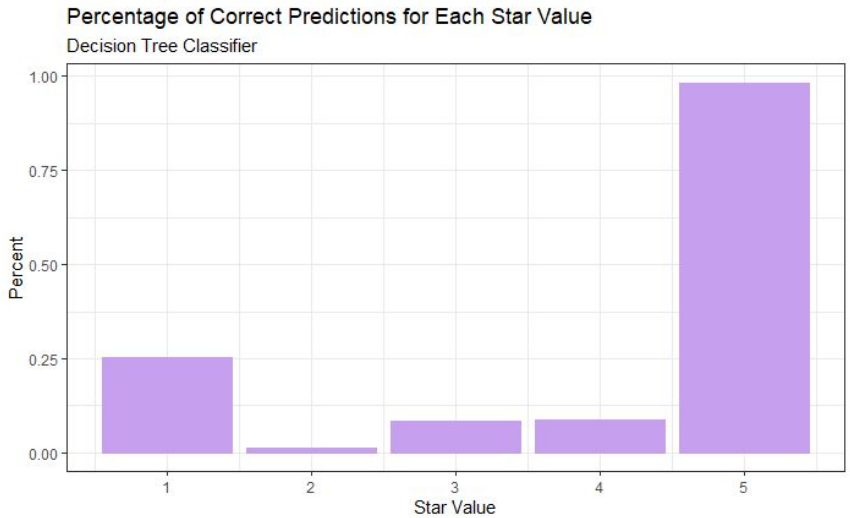

63.1%
Logistic Regression
Regularization Parameter: 0
Max Iterations: 100
Thresholds: None
63.2%
Decision Tree
Max Depth: 5
Minimum Info Gain: 0
Impurity: gini
62.8%
Random Forest
Max Depth: 5
Minimum Info Gain: 0
Impurity: gini

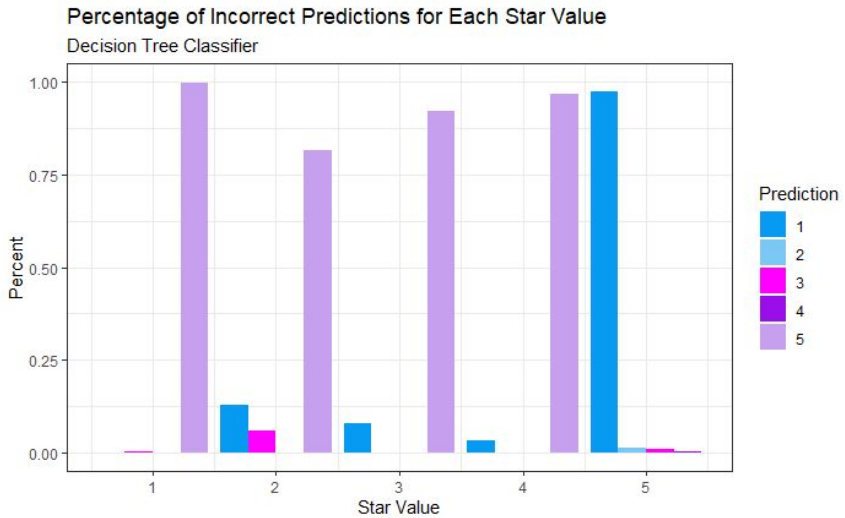
Evaluation - Logistic Regression Classifier



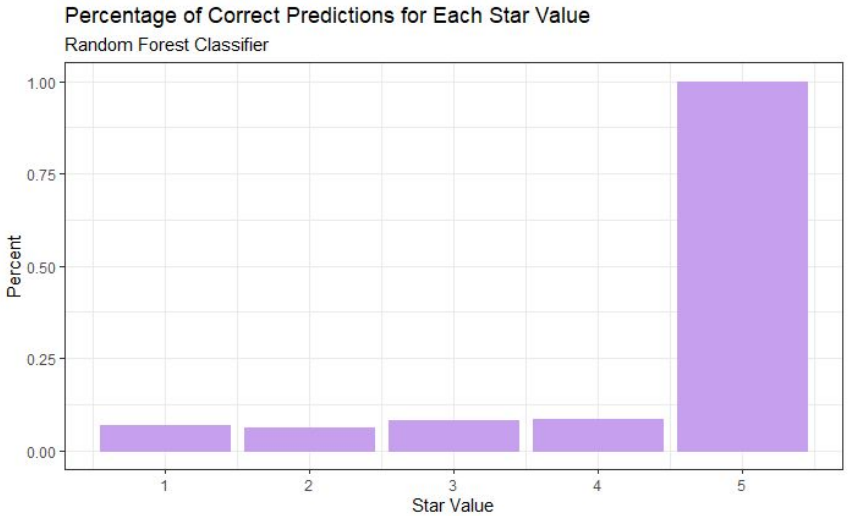


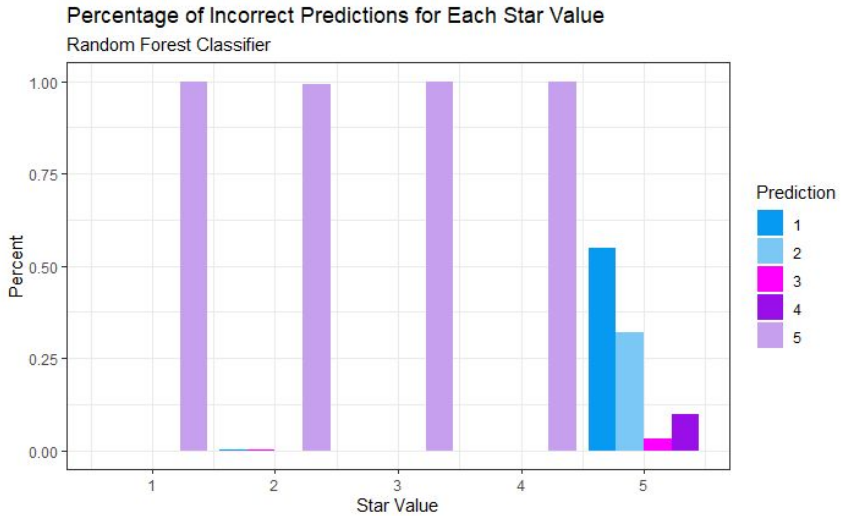
Evaluation - Decision Tree Classifier





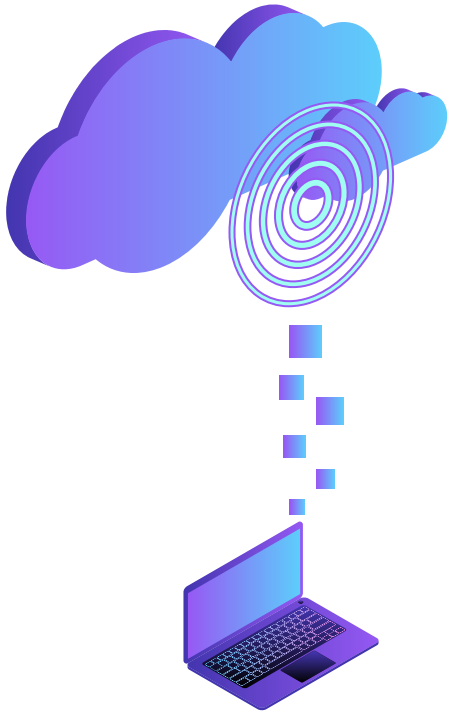
Evaluation - Random Forest Classifier





Challenges

Time

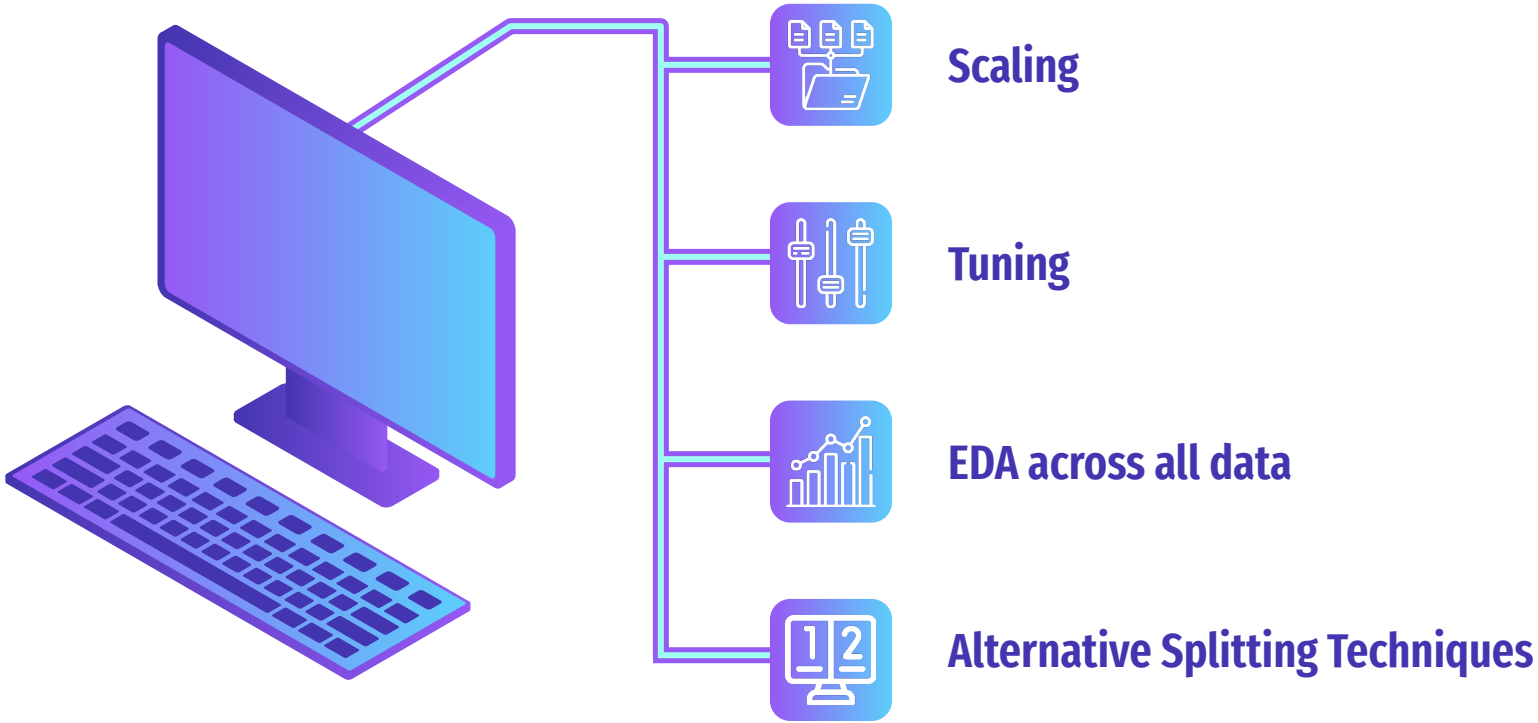


Data Size

Model Availability

Subject Matter Expertise

Next Steps


Scaling
Tuning
EDA across all data
Alternative Splitting Techniques

Questions?