



Transforming the Coffee-Buying Experience

Andrea Ernesto Bianchi
Philip Bachas-Daunert
Will Jaffee
Dylan Mason
Josh Rochlin

Table of Contents

01

Introduction

Fast facts &
motivation

02

EDA

Graphs & plots

03

Model Building

Our methods

04

Evaluation

Model
interpretation &
performance

05

De-brief

Key findings

A close-up photograph of a professional espresso machine. Two light blue ceramic cups are placed on the machine's drip tray. The cup on the right is being filled with a stream of dark coffee from a chrome spout. A pressure gauge is visible on the right side of the machine. The left side of the image is partially covered by a light pink circular graphic containing text.

01

Introduction

Fast facts & motivation

Fast Facts

America runs on...
Coffee!

An economic force
~75% of Americans
drank coffee in 2022



18-29 domination
33% of this demographic
visited a shop in 2020

Adapting with tech
Pre-shop ordering, QR
codes, recipe costing

Our Motivation

How can a coffee shop improve its business?



Approach a target market that will be loyal



Understand who its customers are



Offer great products and a welcoming atmosphere



Identify problems that customers encounter when in store



02

Exploratory Data Analysis

Graphs and plots

Data Columns

transaction_id: unique identifier for each transaction

age: age of the customer making the transaction

income: income of the customer making the transaction

sex: sex of the customer making the transaction

rewards_member: whether or not the customer is a rewards program member

occupation: the occupation of the customer making the transaction

num_items: number of items purchased in the transaction

purchase_method: method of purchase (e.g., cash, credit card)

wait_time: time spent waiting in line before making the transaction in minutes

purchase_amount: total purchase amount of the transaction in USD

store_location: location of the store where the transaction took place

transaction_time: hour when the transaction took place

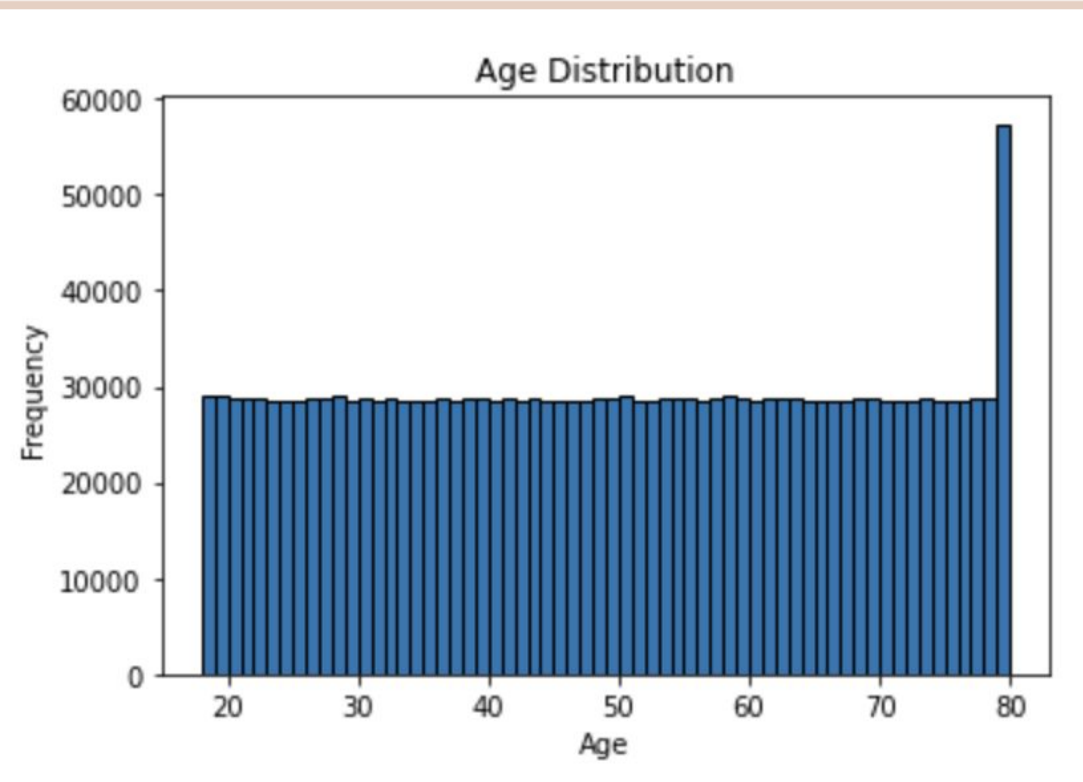
day_of_week: day of the week when the transaction took place

Income Distribution



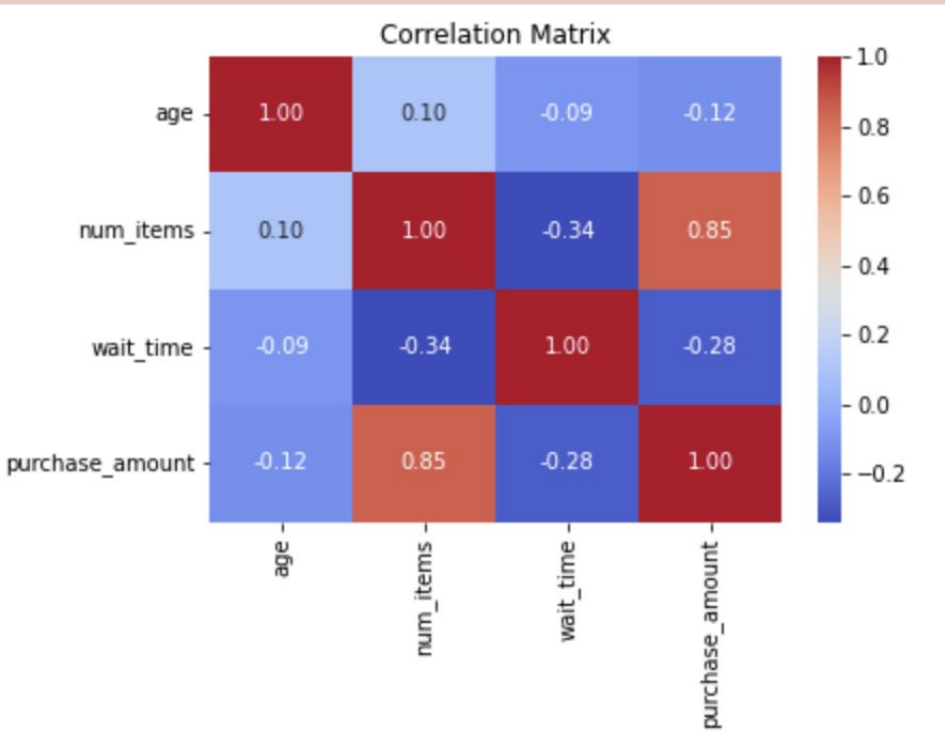
- Most customers are middle-class earners
- Location of shops a factor?
- How does work/job role contribute?

Age Factor



- Heavily left-skewed
- Gen Z domination?
- Buying vs. Hanging Out ?

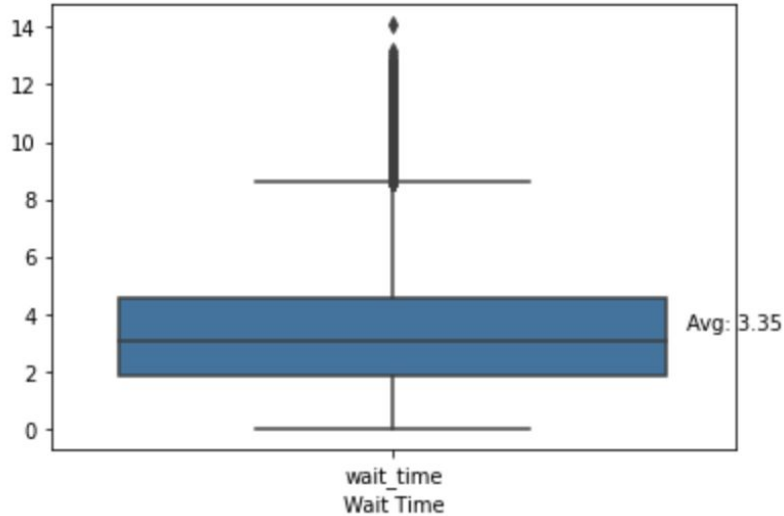
Correlations



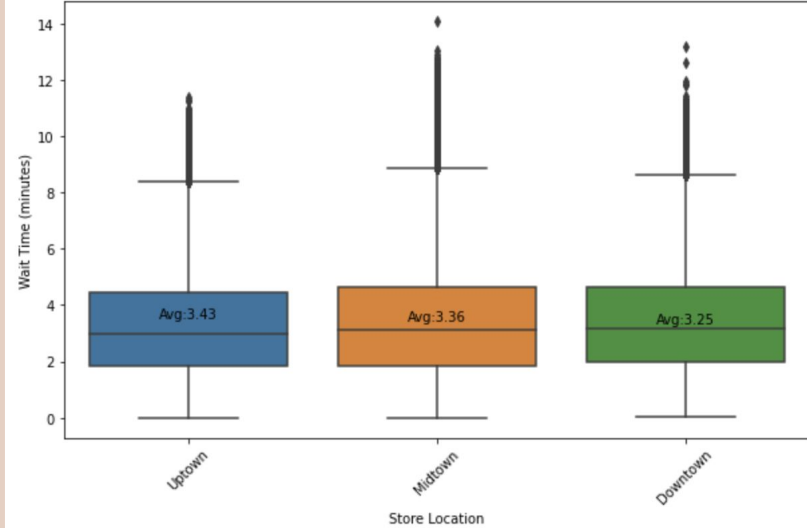
- Buy less, wait less
- Buy more, spend more

Wait Time

Wait Time Distribution

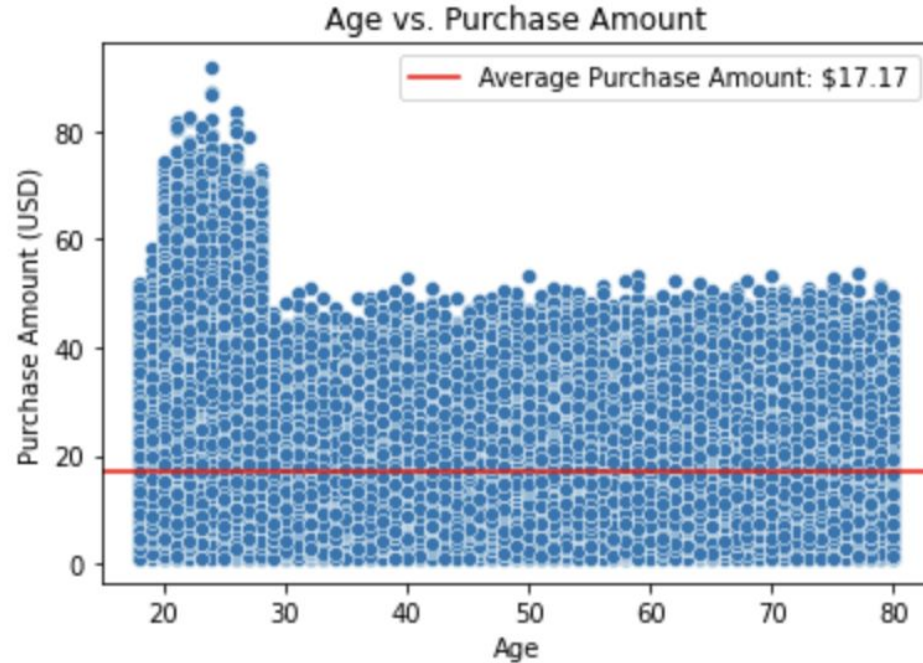


Wait Time Distribution by Store Location



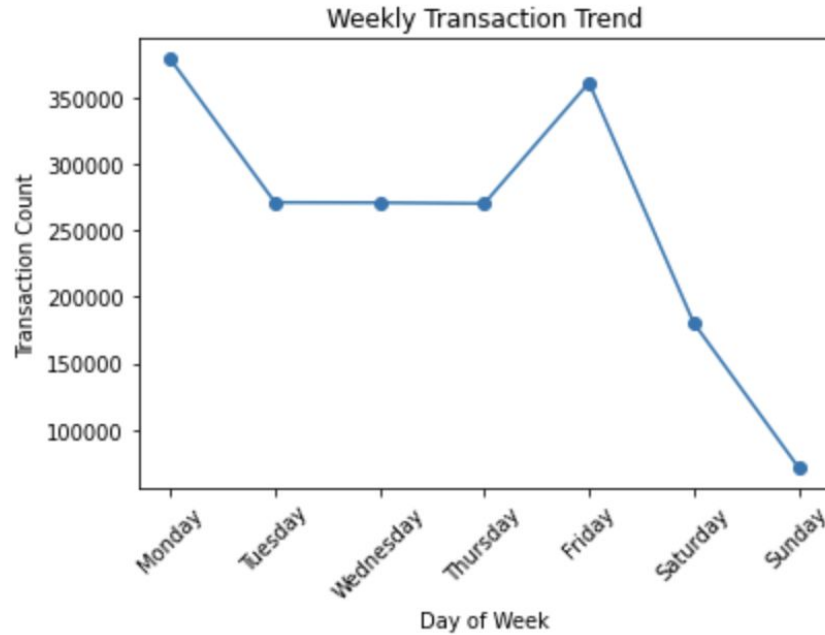
Wait times are consistent throughout different store locations

Purchase Amount by Age



- Average transaction is \$17.17
- Younger customers tend to have more expensive orders than older customers

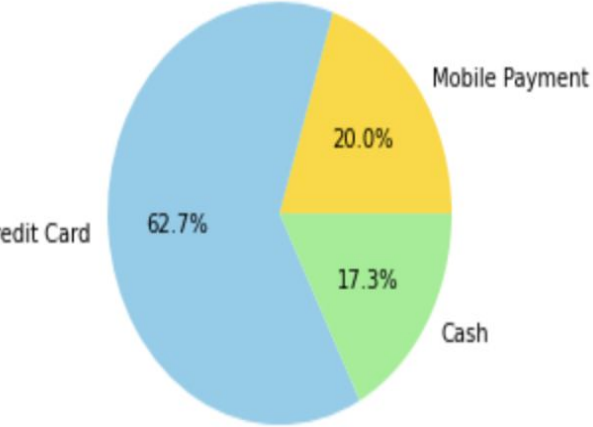
Daily Transaction Trend



- Monday and Friday are most popular days
- Mid-week is busy as well
- Weekends are significantly less busy

Customer Info

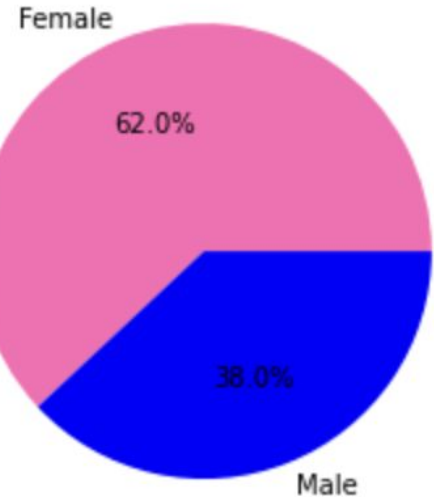
Purchase Method Distribution



Majority of customers pay using non-cash methods (82.7%)

Majority of customers are female

Gender Distribution



The background is a blurred image of a cafe interior. In the foreground, a white cup of coffee with a latte art design sits on a white saucer. Below the saucer, there are some pastries, possibly muffins or breads, on a table. The overall lighting is warm and cozy. There are also some decorative elements: a large light brown circle in the top right corner, a smaller light brown circle in the bottom left corner, and a pattern of small white dots in the bottom left corner.

03

Model Building

Trees, Linear, Logistic

Tree Methods Comparison



Decision Tree

88.1% accuracy

.292 RMSE



Random Forest

88.2% accuracy

.291 RMSE



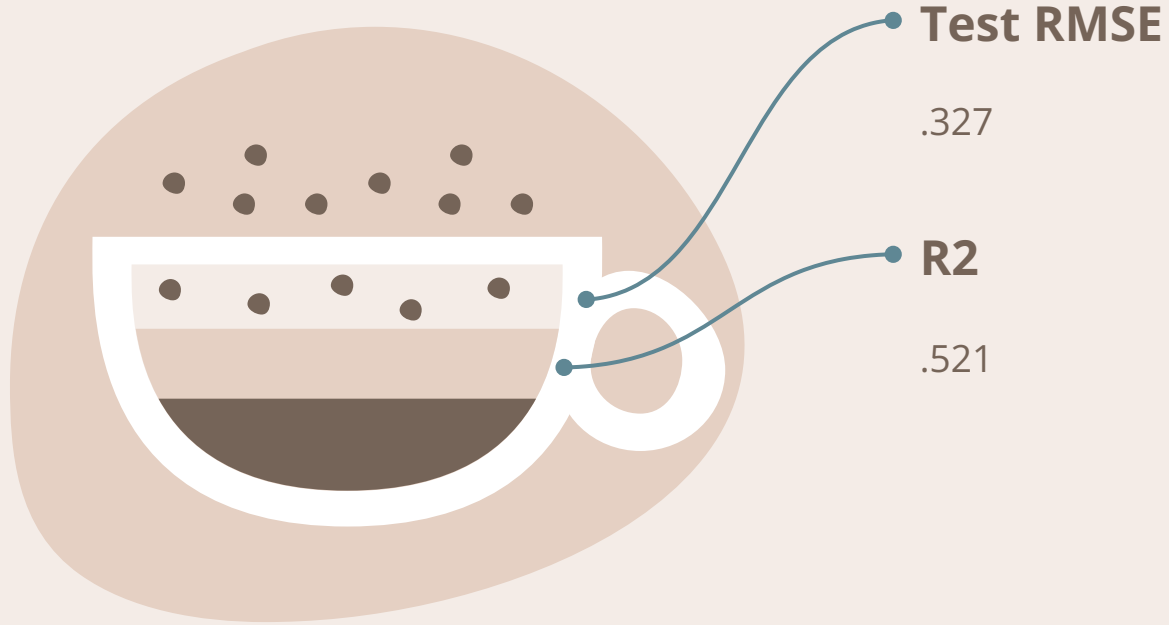
Gradient Boosting

89.6% accuracy

.273 RMSE

Used Features: ['transaction_id', 'age', 'num_items', 'wait_time', 'purchase_amount', 'transaction_time'] to predict if the customer is a rewards member and did a 80,20 train, test split.

Linear Regression



Logistic Regression



Step 1

Used indexer and coder to transform variables into numerical



Step 2

Used assembler to create one vector



Step 3

Created log model to predict rewards member



Step 4

Used pipeline function to expedite pre-processing



Step 5

Reported .867 area under the curve, which is very high



04

Final Model Interpretation

Model interpretation & performance

Why is Gradient Boosting most Successful?

Captures non-linear relationships

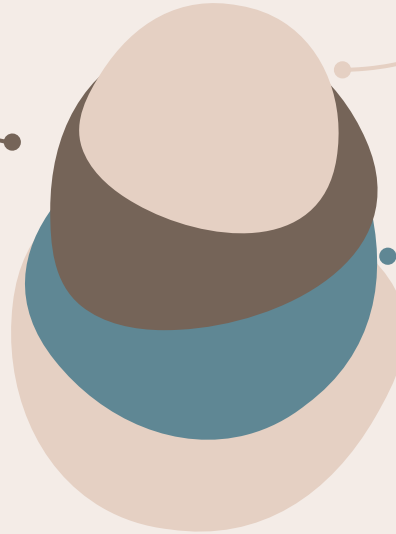
Lots of interactions in the data

Large dataset

Sort through intricate factors effectively

Consistent accuracy

Data is reliable and reflective of shop trends





05

Debrief

Key findings

Most Influential Features for Rewards Program Members

Age

Target customers in the 18-49 demographic

Purchase amount

Visit more, spend more, engage more

Number of items

Loyalty incentives come from buying more items (point accumulation)

Wait time

Shorter wait times = improved satisfaction = potential loyalty

Transaction time

Morning hours during the week are busiest



Thank you for listening

Feel free to ask any questions



References

1. Boyarsky, K. Toast Tab. (n.d.). Coffee Shop Industry Trends and Statistics. Retrieved from <https://pos.toasttab.com/blog/on-the-line/coffee-shop-industry-trends>