IMPROVING VENTRA DATA QUALITY AT PACE
AN ADVANCED DATA SCIENCE MODEL

James Wilson
Senior Service Analyst
Dr. Moyin Li
Project Manager Transportation Analytics

pace
VENTRA SMART CARD

- Ventra smartcard data: Transactions for each passenger trip
  - Calculate ridership by Route for Pace reports and analysis
  - Ventra usage

Ventra smart card data

Ridership

Fare Payment

Regional mobility

Reporting
PROJECT PURPOSE

• Prepping imperfect data
  • Deal with high number of null records
  • Correct errors in route information
  • Automate manual counts from non-IBS (AVL) busses and Special Event counts

• Improve Ventra data quality & optimize reporting process
MULTIPLE DATA SOURCES

- Ventra
- IBS
- GIS files
- Hastus/GTFS
STAGE 1 – DIVISION AND IBS LOOKUPS

- **Reassign**
  - Reassign (combine) Division information

- **Correct**
  - Correct Division information (bus moves)

- **Lookup**
  - Lookup Route information from IBS
STAGE 2 – MACHINE LEARNING AND VALIDATION

- **Impute**
  - Impute missing and incorrect route info using machine learning

- **Validate**
  - Validate current and imputed route info (spatial and temporal)
Stage 1

User Input

+ Ventra
+ Div/ Routes
+ Div/ Driver #s

Logical Sequence

Div Still Unknown?

Y N

Stage 0 output

Route avail?

Y N

+ IBS

Lookup Route

Manual and Special Rides

Stage 1 output
Use machine learning to improve data quality
- K-nearest neighbors (KNN) to find routes

To identify k nearest neighbors
- Compute distance from unknown record to other records.
- Euclidean distance is commonly used

\[
d(p,q) = \sqrt{\sum (p_i - q_i)^2}
\]

Think of it as a similarity between two points among several dimensions:
- Date
- Bus ID
- Driver ID
- Time of transaction
- Geospatial distance: Lat/Lon
SPATIAL VALIDATION TO IDENTIFY INCORRECT ROUTE INFORMATION

- Red dots are Ventra transactions identified as Route 209
- Failed spatial validation
- Sent to KNN model and some re-classified as Route 234
- Others re-classified as other cross routes
- Possible reasons:
  - Bus interlining
  - Incorrect login
  - Maintenance related bus exchange
NORTHWEST

RAW DATA

• 16,284 unknown rides
• 7,118 unknown records
• Unknowns on many routes
• Some are obvious
• Difficult to assign
NORTHWEST

STAGE 1

- 3,580 unknown rides
- 2,239 unknown records
- Unknowns still on many routes
NORTHWEST

STAGE 2

• 51 unknown rides
• 33 unknown records
• Very few false positives
RAW DATA

- 12,828 unknown rides
- 5,472 unknown records
- March 15 has many nulls
NORTHWEST WEEKDAYS

STAGE 1

• 3,128 unknown rides
• 1,902 unknown records
• March 15 still high nulls
• Other days look good
NORTHWEST WEEKDAYS

STAGE 2

- 12 unknown rides
- 8 unknown records
- March 15 looks good
RIVER WEEKDAYS

RAW DATA
6,374 unknown rides
2,776 unknown records

STAGE 1
479 unknown rides
277 unknown records

STAGE 2
30 unknown rides
21 unknown records
WEST WEEKDAYS

RAW DATA
- 39,565 unknown rides
- 15,477 unknown records

STAGE 1
- 4,255 unknown rides
- 2,628 unknown records

STAGE 2
- 37 unknown rides
- 14 unknown records
NORTH WEEKDAYS

RAW DATA
8,555 unknown rides
4,030 unknown records

STAGE 1
3,706 unknown rides
2,356 unknown records

STAGE 2
89 unknown rides
56 unknown records
RESULTS

• Analyst time required reduced from 48 to 12 hours per month
• More accurate imputation of nulls
• Route info validated
• Manual and Special Events counts automated

<table>
<thead>
<tr>
<th>Pace System Totals for March 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Unknown</td>
</tr>
<tr>
<td>Rides</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Raw Data</td>
</tr>
<tr>
<td>After Stage1</td>
</tr>
<tr>
<td>After Stage2</td>
</tr>
</tbody>
</table>
TAKEAWAYS

• Data is not perfect. Imperfect data leads to inaccurate decision making

• Other transit agencies can replicate this method to improve data accuracy

• This method demonstrates how multiple data sources can be combined and leveraged for better data-driven decision-making
THANK YOU!

Q & A
KNN VS RF
ROUTE 301

- kNN
  - more accurate
  - more transactions identified

- Random Forest
  - less accurate
  - missing transactions
KNN VS RF
ROUTE 303

- kNN
  - more accurate
  - more transactions identified

- Random Forest
  - less accurate
  - missing transactions