Taxi Pickup Location Recommender System

Strategies for Taxi Drivers to Maximize Revenue

Dooinn KIM

Feb 2024

Agenda

"What do we cover?"

1.Business overview

2.Segment Analysis

3.Recommender System Introduction

4.Development Process

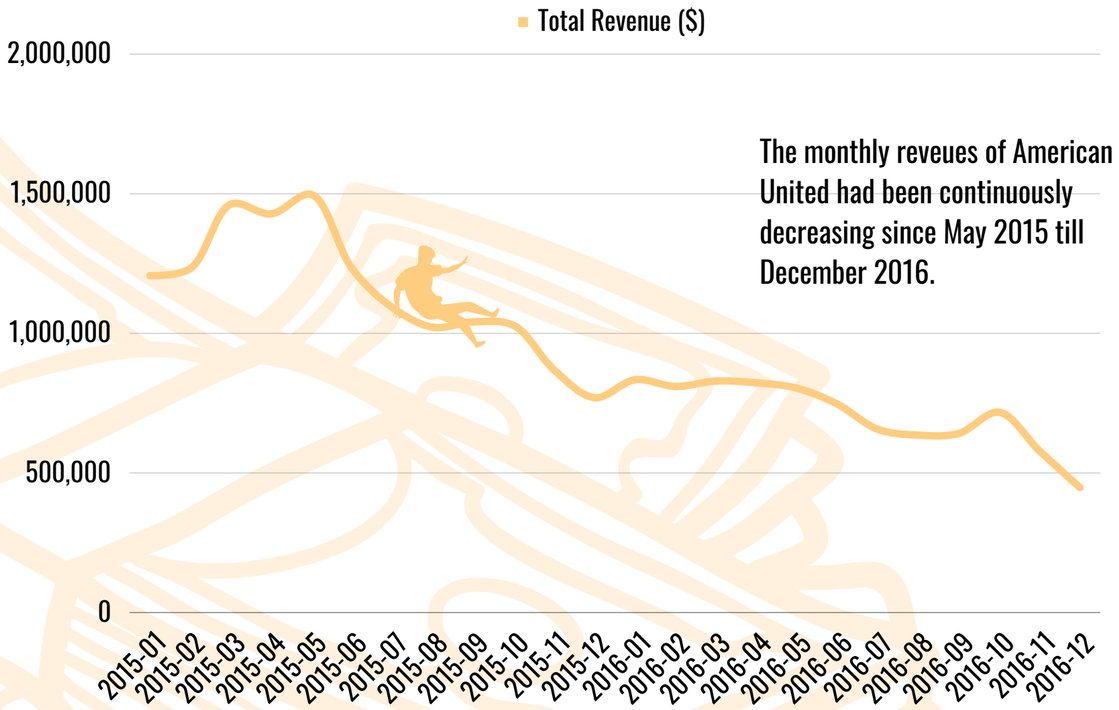
5.Data & Data Sources

6.Conclusion

Business Overview

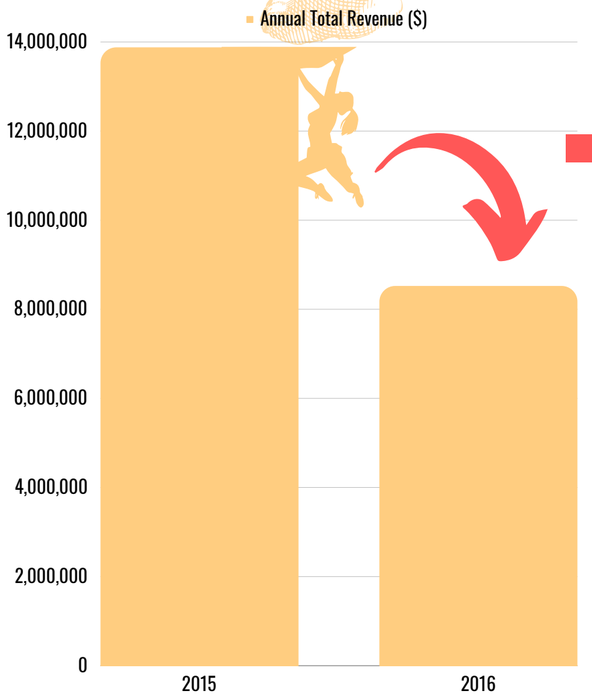
Understanding American United's Current Challenges

Monthly Revenue Trend Jan 2015 - Dec 2016



The monthly reveues of American United had been continuously decreasing since May 2015 till December 2016.

Annual Growth Rate in 2015-2016 (%)



-41%

Total Annual Revenue Comparison (2015 vs 2016)

- 2015: $ 140,343
- 2016: $ 83, 073

Segment Analysis: Earning Grades

Exploring Differences Across Taxi Earning Grades

AVERAGE ANNUAL TOTAL REVENUE BY EARNING GRADES*

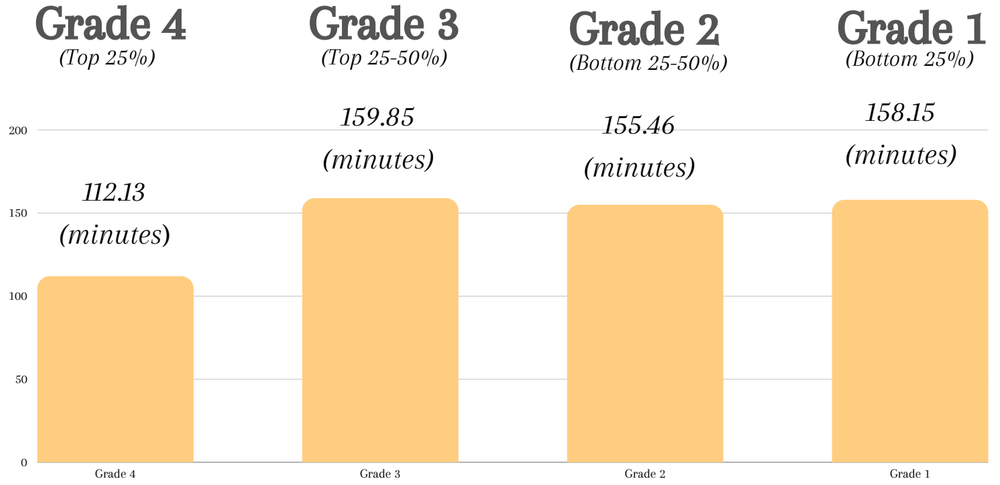
"More trips & longer distance , more money...!"

	Grade 4 *(Top 25%)*	Grade 3 *(Top 25-50%)*	Grade 2 *(Bottom 25-50%)*	Grade 1 *(Bottom 25%)*
Total Earnings: ***(median)***	*$ 89,362*	*$ 64,326*	*$ 39,648*	*$ 23,749*
Total trips ***(median)***	*6,723 trips*	*5,034 trips*	*3,076 trips*	*1,638trips*
Total Distance ***(median)***	*25,428 miles*	*17,974 miles*	*11,099 miles*	*6,571 miles*

***Earning Grade:** This is a grading system that divides annual total earnings per taxi into quartiles. Grade 4 represents the top 25% of earners (>75th percentile), Grade 3 includes earners from the 50th to the 75th percentile (top 25-50%), Grade 2 covers earners from the 25th to the 50th percentile (bottom 25-50%), and Grade 1 represents the bottom 25% of earners (<25th percentile).*

AVERAGE TURNAROUND TIMES** BY EARNING GRADES

"Taxis in Grade 4 have Average 45 minutes less Turnaround times!"



Comparion of the length of turnaround times by different grades.

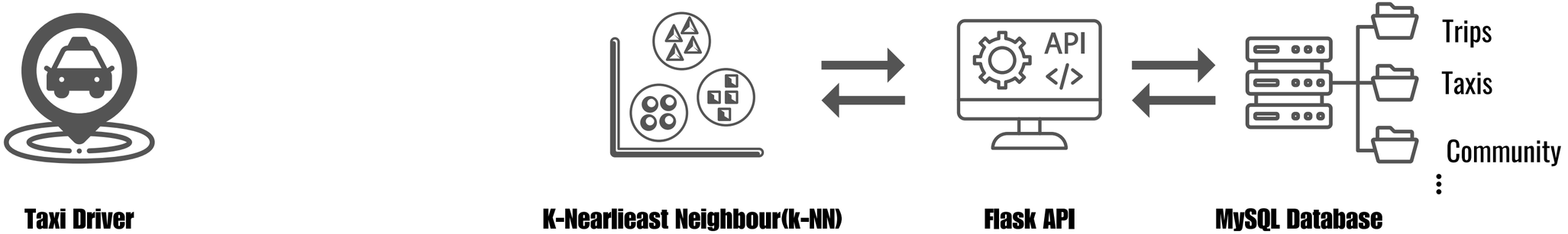
***Turnaround Times:** This refers to the duration between the last drop-off time and the next pickup time for a taxi. Turnaround times are calculated by subtracting the drop-off timestamp from the next pickup timestamp for each taxi, and then aggregating these durations to determine an average.*

Location Recommender System

How the System Operates

Input

- Current Location (Latitude & Longitutde)
- Current Day & Hour



Taxi Driver → **K-Nearlieast Neighbour(k-NN)** → **Flask API** → **MySQL Database**

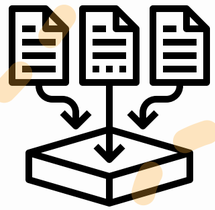
Output

- Recommendation of top 5 nearleast & most popular pickup locations

Development Process

Journey from Data Collection to ML Implementation

1. Data Collection

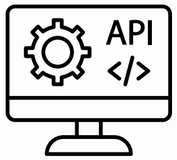


- **BIg Data system**: Big Query Google Cloud Platform Chicago Taxi Trips
- **Webscraping**: Wikipedia - Chicago Community Areas
- **API**: Nomiatim API - OpenStreetMapa Geodata
- **Flat Files:** Chicago Data Portal - City of Chicago

2. Database Building

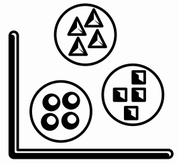
- **Database:** MySQL
- **Schema:** ‘chicago_taxi’
- **Tables:** ‘trips’, ‘taxi’, ‘company’, ‘location’, ‘comuunity’

3. API Creation



- **API:** Flask API
- **Data expose: ‘**GET/ community’, ‘GET/ trips’, ‘GET/ company’, ‘GET/ location’, ‘GET/ trips’

4. ML Devlopment



- Optimal Cluster Determination - Silhouette Scores
- Clustering with K-means
- K-Nearest Neighbors (K-NN) for cluster center prediction
- Intergration with Nominatim API and Vizualization with Folium maps

Data & Data Sources

Comprehensive Data Integration

Data Sources

Source1: Big Data System

Big Query Google Cloud Platform - Chicago Taxi Trips

chicago_taxi *(4,972,163 rows x 24 columns)*

Columns	Description	Data Type
unique_key	Unique identifier for each record	object
taxi_id	Unique identifier for each taxi	object
trip_year	The year of trips made (e.g 2015)	Int64
trip_start_timestamp	Timestamp marked of trip start	datetime64[us, UTC]
trip_start_date	Date of a trip start (e.g. 2015-04-04)	dbdate
trip_start_time	Hours of a trip made (e.g. 15:00:00)	object
trip_end_timestamp	Timestamp marked of trip end	datetime64[us, UTC]
trip_end_date	Date of a trip end (e.g. 2015-04-04)	dbdate
trip_end_time	Hours of a trip made (e.g. 15:00:00)	object
trip_seconds	Total duration of a trip in seconds	Int64
trip_miles	Total distance of a trip in miles	float64
pickup_community_area	Pickup location of Chicago community area number	Int64
dropoff_community_area	Dropoff location of Chicago community area number	Int64
fare	Taxi fare by distance	Int64
tips	Tips given by passengers	Int64
tolls	Tolls for highways	Int64
extra	Extra charges by taxi drivers	Int64
trip_total	Total amount of trip (fare + tips + tolls + extra)	Int64
payment_type	Passenger's payment method - Card/Cash	object
company	Taxi companies in Chicago	object
pickup_latitude	-	float64
pickup_longitude	-	float64
dropoff_lattitude	-	float64
dropoff_longitude	-	float64

Source2: Web scraping

Wikipedia - Chicago Community Areas

community *(77 rows x 3 columns)*

Columns	Description	Data Type
community_number	Chicago Community Area number	Int64
communit_name	Name of Chicago Community	object
Population	Population of community	Int64

Source3: API

Nominatim API - OpenStreetMap Data

location_info *(421 rows x 3 columns)*

Columns	Description	Data Type
location_coordinates	Location coordinates	object
address	Address of location coordinates	object
type	Type of building	object

Source4: Flatfiles

Chicago Data Portal- Taxi Vehicle Model & Make

taxi *(386 rows x 6 columns)*

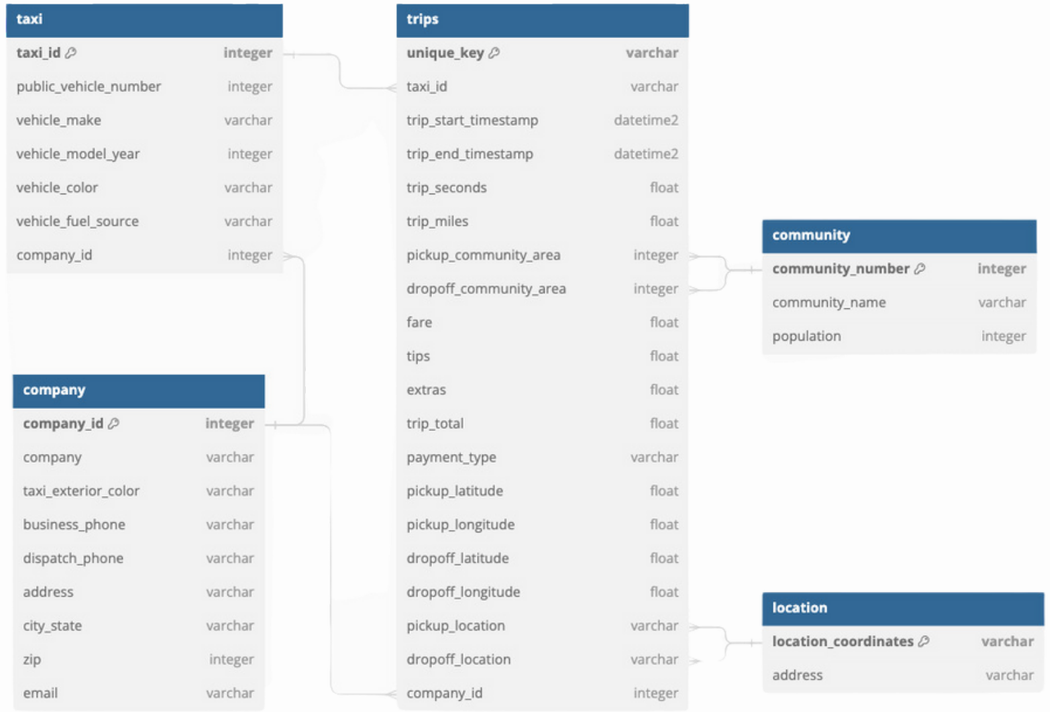
Column	Description	Data type
taxi_id	Unique identifier for each taxi	object
Public Vehicle Number	Taxi Licence identifier	int64
Vehicle Make	Name of vehicle brand/company (e.g. Ford, Kia)	object
Vehicle Model Year	Model made year of the vehicle	float64
Vehicle Color	Color of vehicle	object
Vehicle Fuel Source	Fuel source type (e.g. Hybrid, Flex Fuel)	object

Chicago Data Portal - Chicago Taxi Company

company *(19 rows x 9 columns)*

Column	Description	Data type
company_id	Unique company identifier of taxi company	int64
company	Name of taxi company	object
taxi_exterior_color	-	object
business_phone	-	object
dispatch_phone	-	object
address	Company Address	object
city_state	State where company located	object
zip	Zipcode of company location	into64
email	-	object

Entity Relationship Diagram (ERD)



Conclusion

Challenges, Limitations, and Next Steps

Limitations & Challenges

- High volume of Data
- Limited Availability of Relevant Data
- Machine Learning Performance Evaluation
- Training Dataset Geographical Limitation

Next Steps

- Expanding the dataset beyond Chicago taxi trips
- Training data from years beyond 2015 and a borader selection of companies
- Experimenting with and evaluating the recommender system in real-world scenario.
- Real-time and competitor analysis