System Insights from Uber Ride Data

A Data-Driven Analysis

Author: Marko Mirković

Presentation outline

1. System & data overview
2. Key analyses & insights
   - Vehicle performance
   - Location hotspots
   - Hourly ride trends
   - Long ride vehicle choises
   - Cancellation causes
   - Payment methods
3. Predicting cancellations
4. Summary of recomendations
5. Q&A

Understanding the system

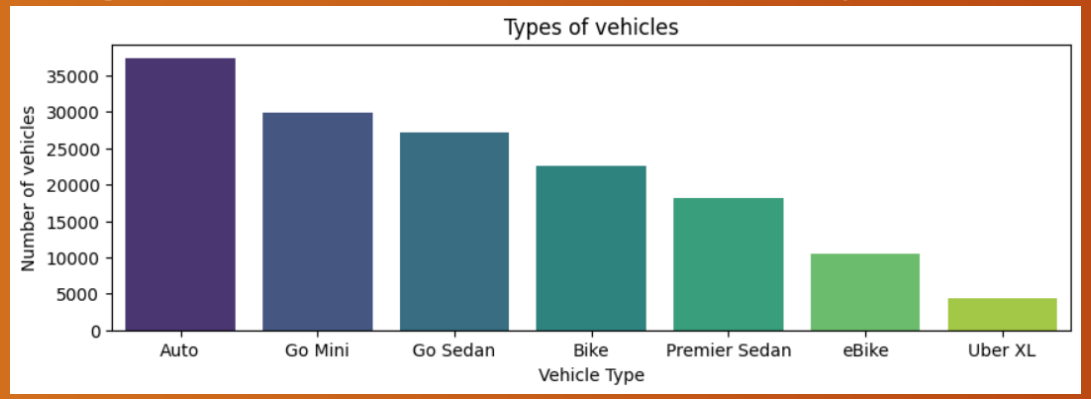
- Domain: On-demand ride-hailing service.
- Goal: Analyze the Uber system to discover insights that can improve operational efficiency and revenue.
- Stakeholders:
  - Uber operations: Optimize driver availability and vehicle allocation.
  - Customers: Improve ride experience and reliability.
  - Drivers: Maximize earnings and efficiency.

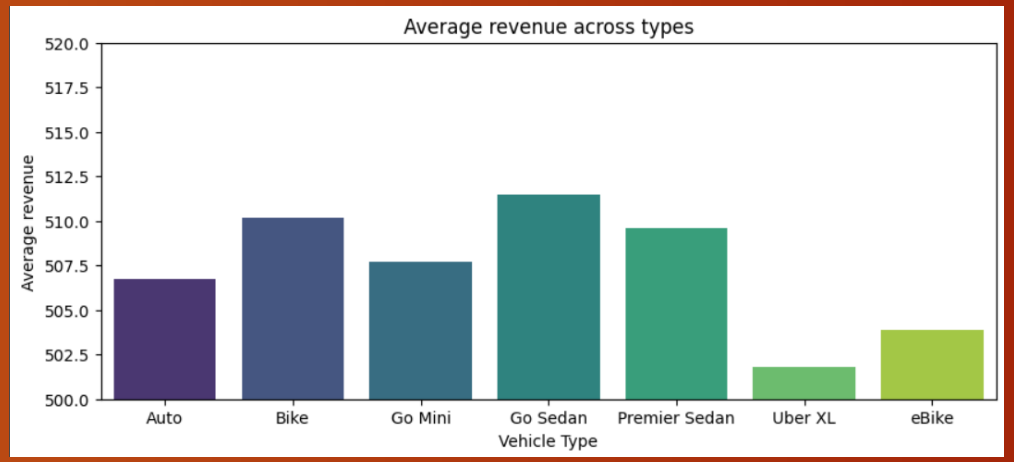
Dataset: Uber Data Analytics

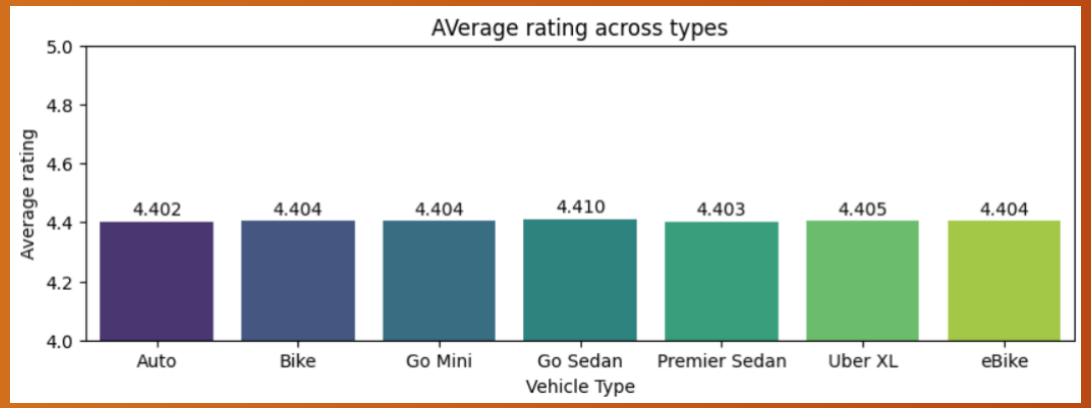
- Source: Kaggle - Uber Data Analytics Dataset
- Description: ~150k instances in ride bookings with 21 features.
- Key features: Booking Value, Ride Distance, Driver Ratings, Customer Rating, Vehicle Type, Pickup/Drop Location ...
- Data Prep: Converted date/time, handled missing values for analysis, and created an Hour feature for trend analysis.

Analysis: Which Vehicle Type Performs Best?

- Question/Hypothesis: Which vehicle type is most common, highest-rated, and generates the most revenue per ride?



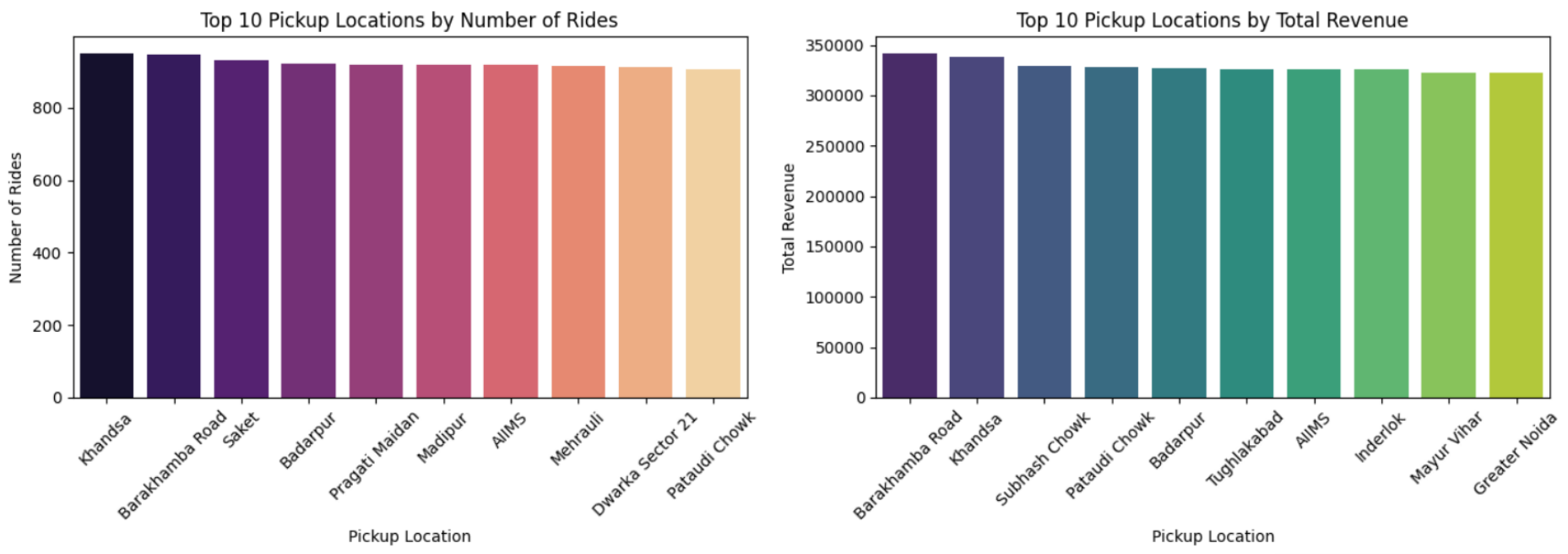




- Auto is the most frequent vehicle type.
- Customer ratings are nearly identical across all types.
- Go Sedan generates the highest average revenue per ride.
- To maximize revenue, Uber should encourage drivers to invest in Go Sedans or increase their presence in the fleet.

Analysis: Where Are the Hotspots for Rides & Revenue?

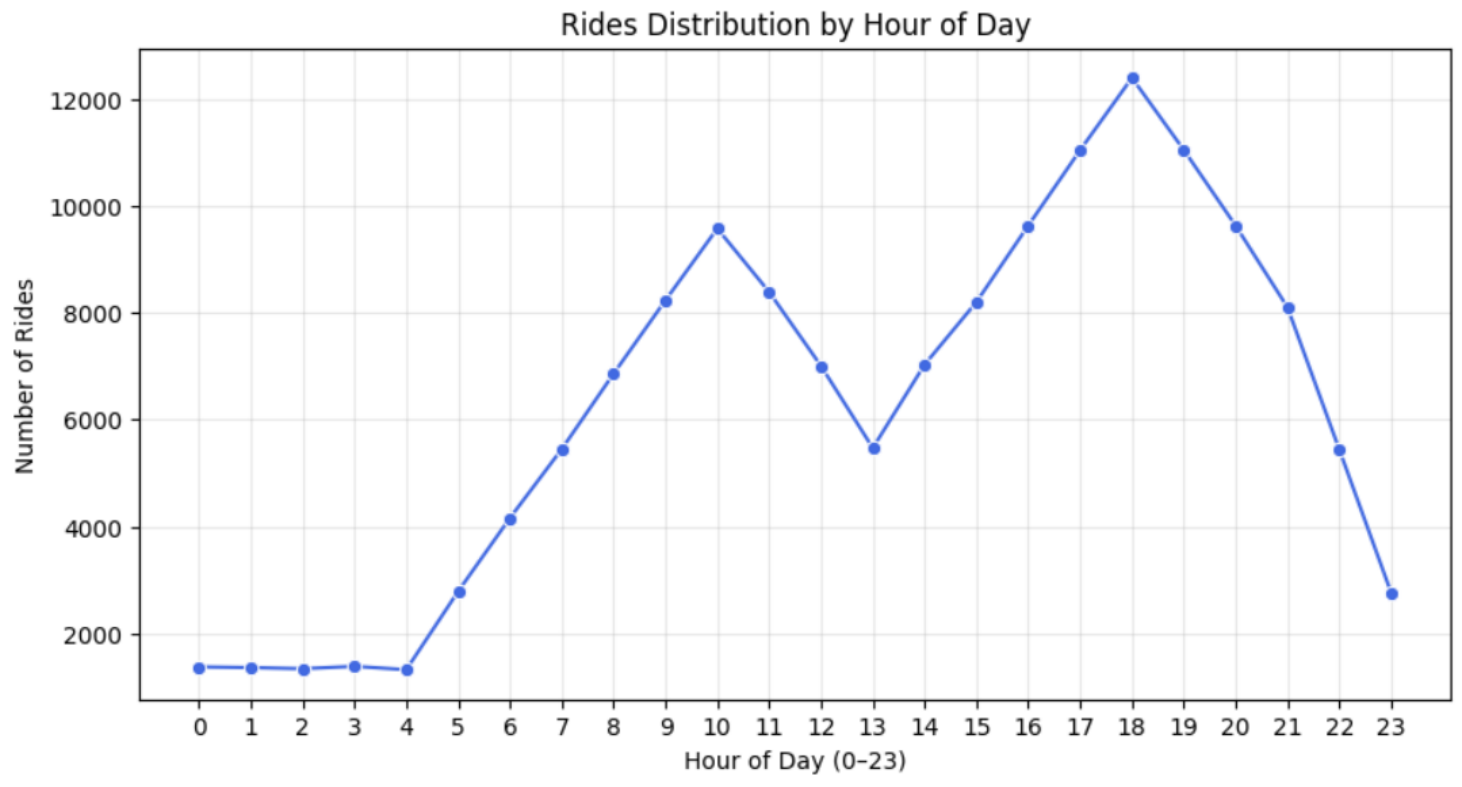
- Question/Hypothesis: Which pickup locations have the highest frequency of rides and generate the most revenue?



- Locations like Khandsa, Barakhamba Road, Badarput appear in the both top chars. Increase vehicle coverage and driver availability in these key hotspot locations to reduce wait times and capture more bookings, especially during peak hours.

Analysis: When Are Rides Most In-Demand?

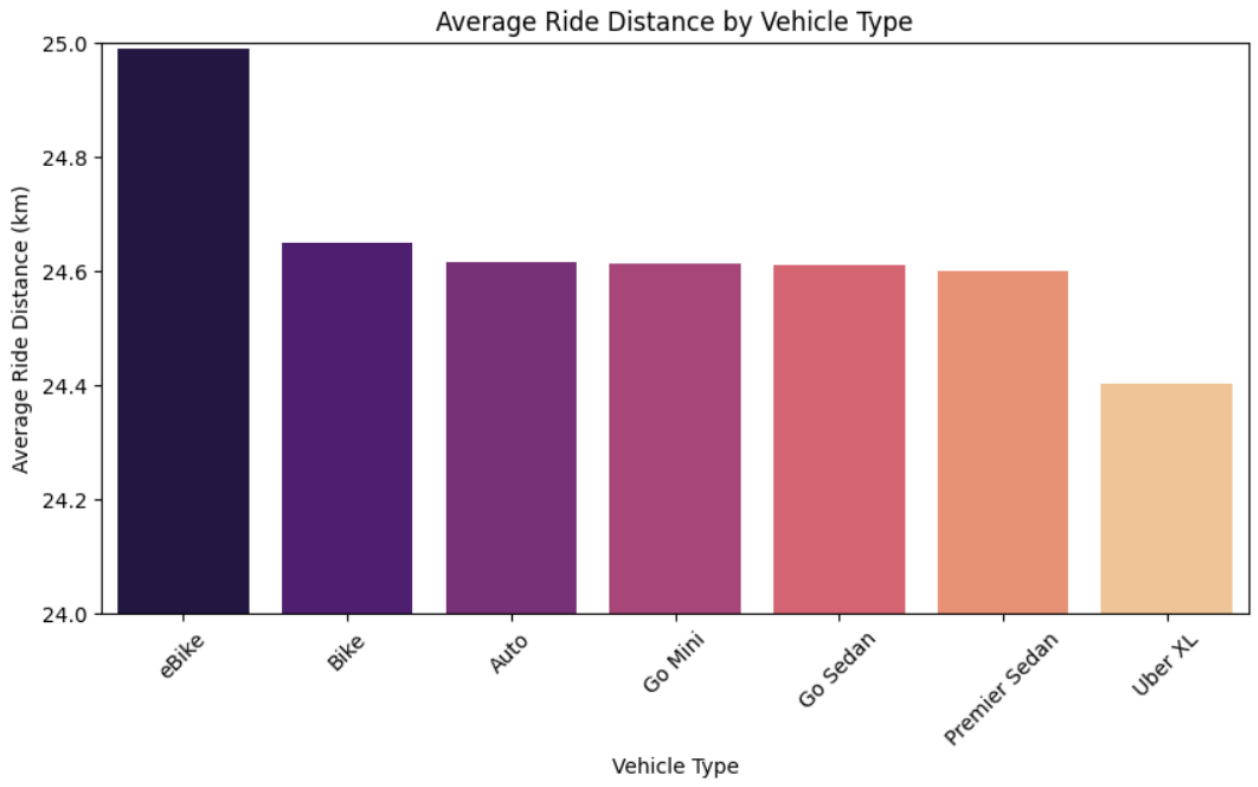
- Question/Hypothesis: What are the peak and off-peak hours for ride bookings throughout the day?



- There are two distinct peaks: a morning commute peak and a larger evening peak.
- The lowest demand is in the early morning hours (around 4 AM).
- Implement a dynamic pricing model. Offer incentives for drivers during peak hours and consider lower promotional prices during off-peak times to stimulate demand.

Analysis: Are Certain Vehicles Used for Longer Rides?

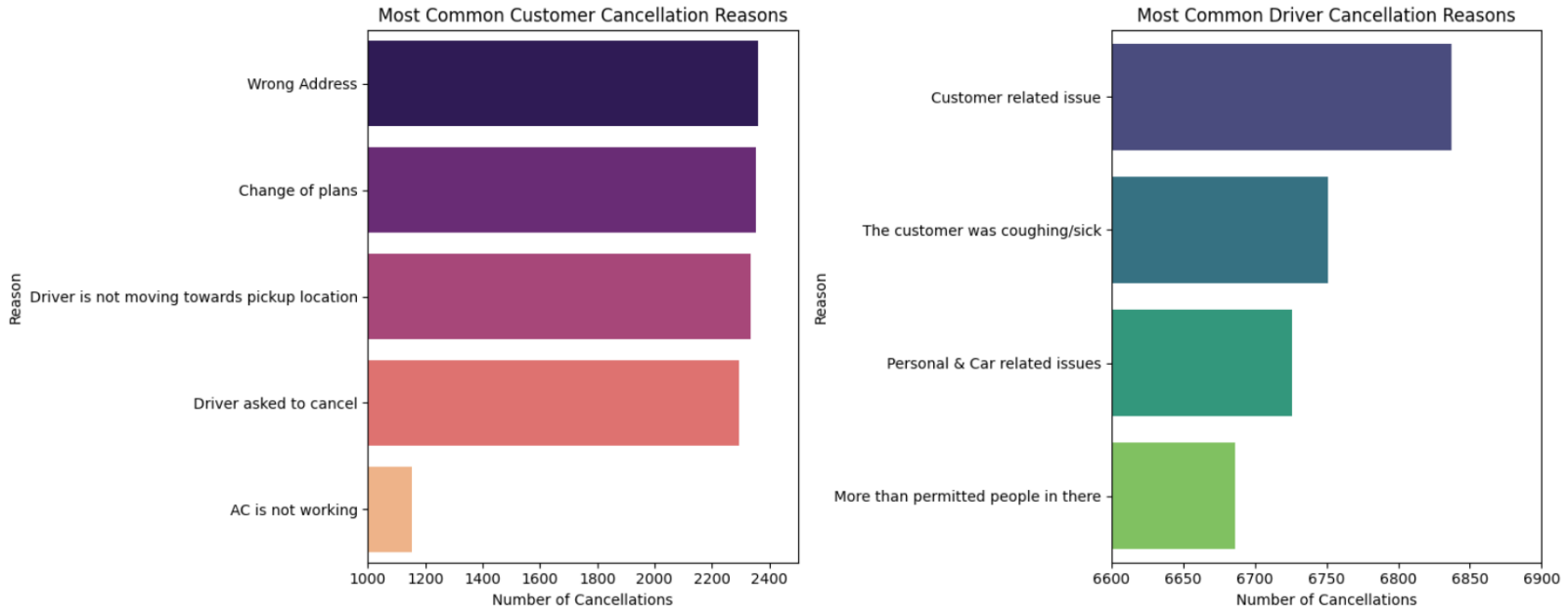
- Question/Hypothesis: Are specific vehicle types more commonly used for longer distances?



- All vehicle types have a very similar average ride distance, with no statistically significant difference.
- This hypothesis is invalidated. We can conclude that vehicle type is not a factor in the distance of a ride. This is a useful finding as it prevents investing in distance-based strategies for specific vehicle types.

Analysis: Why Are Rides Cancelled?

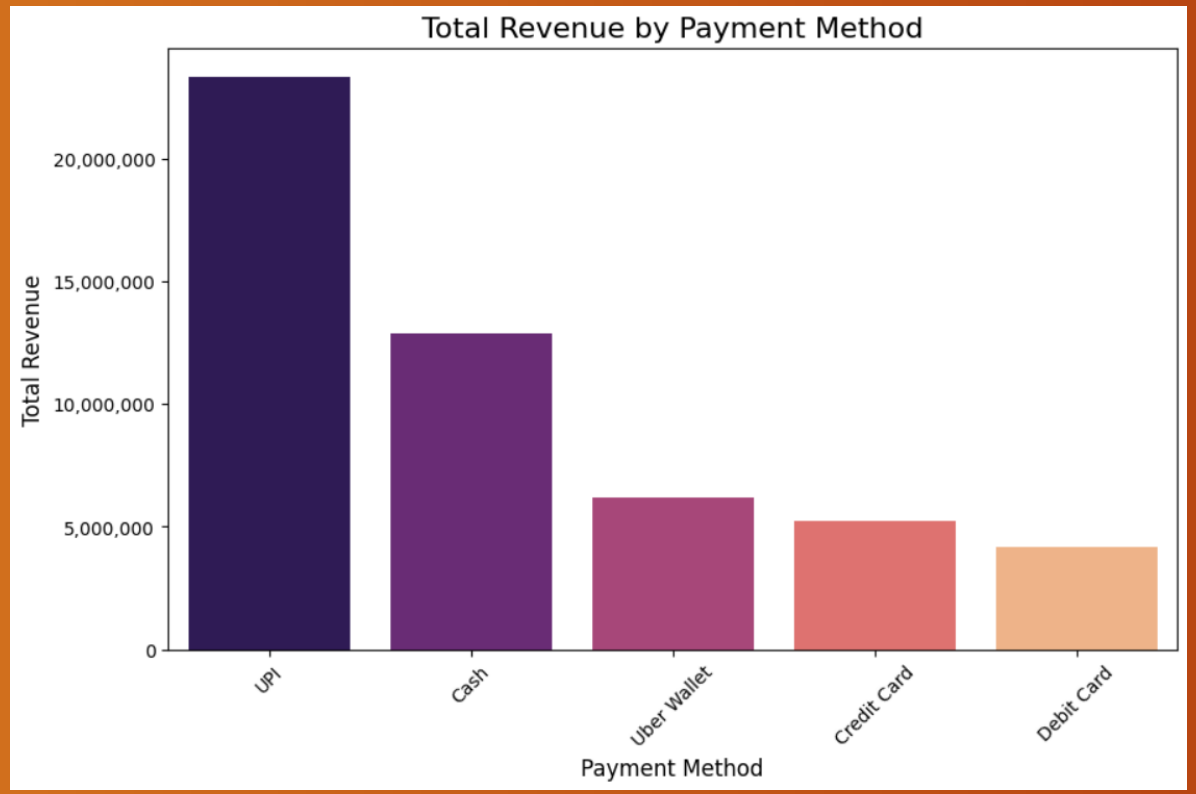
- Question/Hypothesis: What are the most common reasons for cancellations by both customers and drivers?



- Customers: Primarily cancel because the driver is not moving or denies the destination.
- Drivers: Overwhelmingly cancel due to customer related issues.
- Address navigation and communication issues. Potential solutions include improving in-app navigation accuracy.

Analysis: Which Payment Method is Most Valuable?

- Question/Hypothesis: Which payment method is most common and brings in the most revenue?



- UPI is the clear leader, generating the highest total revenue. Cash is also a very popular method.
- Create special promotions or loyalty rewards for customers who use UPI to further encourage its adoption and streamline the payment process.

Advanced Analysis: Predicting Ride Cancellations

- Question/Hypothesis: Can we accurately predict if a ride is likely to be canceled?
- Developed a machine learning pipeline to classify rides as "Cancelled" or "Not Cancelled".
- Compared three models: Logistic Regression, Random Forest, and XGBoost.

	model	train_accuracy	test_accuracy	train_precision	test_precision	train_recall	test_recall	train_f1	test_f1
0	RandomForest	0.999983	0.997433	1.000000	1.000000	0.999933	0.989733	0.999967	0.994840
1	XGBoost	0.997467	0.997433	1.000000	1.000000	0.989867	0.989733	0.994908	0.994840
2	LogisticRegression	0.930125	0.929500	0.781555	0.780031	1.000000	1.000000	0.877385	0.876424

- The Random Forest and XGBoost models achieved ~99.7% accuracy on the test set.
- This high-precision model can be integrated into the booking system. If a ride is flagged with a high probability of cancellation, the system could proactively assign a backup driver, ensuring a more reliable service for the customer.

Key Recommendations for Uber

- Fleet Strategy: Incentivize the use of Go Sedans to increase average revenue per trip.
- Operations: Increase driver density in identified hotspot locations.
- Pricing: Implement dynamic pricing to capitalize on peak demand and encourage off-peak usage.
- Product: Improve in-app navigation and communication to reduce cancellations from both sides.
- Marketing: Launch promotions for customers using the UPI payment method.
- Technology: Deploy the cancellation prediction model to proactively re-assign drivers and improve service reliability.

Thank you!

Q&A