Didi Business Intelligence Challenge

Michelle Sandoval · Jun 2025

🗂️ Project structure:

data/ – raw & processed data

notebooks/ – notebooks EDA, SQL query validation, forecasting

sql/ – .txt files with SQL queries

figures/ – output plots from forecasting model

slides/ – this presentation

README.md – overview & setup

requirements.txt

(I added genre_name and area_name for business context)

📊 SQL Queries & Insights

1. Top-5 Holiday Restaurants

Query: Identify the top 5 restaurants with the highest average number of visitors on holidays, with average per restaurant.

	restaurant_id	genre_name	area_name	avg_holiday_visitors
0	db80363d35f10926	Dining bar	Hokkaidō Asahikawa-shi 6 Jōdōri	7.275000
1	bb09595bab7d5cfb	Izakaya	Niigata-ken Niigata-shi Teraohigashi	5.833333
2	e053c561f32acc28	Izakaya	Hokkaidō Asahikawa-shi 6 Jōdōri	5.240000
3	24b9b2a020826ede	Japanese food	Fukuoka-ken Kitakyūshū-shi Ōtemachi	4.333333
4	42c9aa6d617c5057	Italian/French	Hyōgo-ken Kakogawa-shi Kakogawachō Kitazaike	4.228571

✔ **Insight:** Izakaya genre dominates holiday demand, claiming 2 of the top 3 spots. The top-performing Dining Bar in Asahikawa (Hokkaidō) has 25% higher holiday traffic than its closest competitor, suggesting strong regional appeal for social dining during celebrations.

2. Busiest Day of the Week

Query: Determine which day of the week has the highest average number of visitors.

	day_of_week	avg_visitors	total_records
0	Friday	4.454754	1746
1	Wednesday	4.216495	873
2	Thursday	4.115640	1055
3	Monday	4.049140	814
4	Saturday	3.983149	2077
5	Tuesday	3.913649	718
6	Sunday	3.492447	993

✔ **Insight:** Friday is the clear peak dining day (4.45 avg visitors), while Sunday sees the lowest traffic (3.49) – a 28% drop. This reveals a strong 'pre-weekend' dining culture over actual weekend days.

3. Week-Over-Week Visitor Growth

Query: Calculate weekly growth percentage over the last 4 weeks.

	week_start	total_visitors	prev_week_visitors	growth_pct
0	2017-05-15	78.0	170.0	-54.12
1	2017-05-08	170.0	130.0	30.77
2	2017-05-01	130.0	618.0	-78.96
3	2017-04-24	618.0	1469.0	-57.93

✔️ **Insight:** Visitor counts show extreme volatility, with 3 of 4 weeks seeing >50% declines. The only growth week (+31%) was immediately followed by a 54% crash, indicating unstable demand or data anomalies.

4. 🧩 Six-Month Forecast – LSTM Model

Approach

- Univariate LSTM on daily aggregated visitors
- 90-day input window → next-day prediction
- Train / test split: 80 % / 20 % (≈ last 6 months as hold-out)
- 50 epochs, batch 32, Adam optimiser

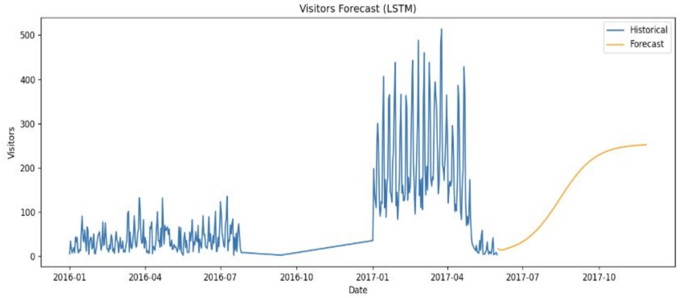
Validation (hold-out)

Metric	Value
MAE	72 visitors
RMSE	92.7
MAPE	232.5%

> Interpretation — MAE ≈ 72 means the model is off by ~72 visitors/day on average.
> MAPE is inflated because many days have < 50 visitors; 1-day errors look large in %.

Shows convergence by epoch ≈ 40.

Historical + 6-Month Forecast



Data period: Jan 2016 – May 2017 | Model trained in Python 3.9 / TensorFlow 2.19

📈 Growth Strategy

5. Double Visitors in 6 months

Based on the data and forecast, here’s a focused 3-phase strategy to double visitors in 6 months:

Phase 1: Stabilize Volatility (Month 1-2)

Problem: 78% weekly visitor swings hurt capacity planning.

1. Dynamic Pricing Engine:

   - Algorithm: IF day = Tuesday/Wednesday → 20% off premium menu

   - Data hook: Use the 28% Sunday-Friday gap to incentivize off-peak days.

2. Demand-Smoothing Partnerships:

   - Partner with theaters/cinemas: "Dinner + Show" bundles on low-traffic Sundays.

   - Technical integration: Sync reservation APIs for real-time inventory.

Target: Reduce weekly volatility by 40%, boost Tuesday traffic 25%.

5. Double Visitors in 6 months

Phase 2: Attack the Plateau (Month 3-4)

Problem: Forecast shows hard ceiling at 250 visitors/day.

1. Predictive Menu Optimization:
   - Technique: Time-series clustering of order data × visitor patterns.
   - Action: Highlight high-margin dishes on forecasted low-traffic days.
2. Genre Expansion:
   - Capitalize on #1 insight: *Izakayas dominate holidays*.
   - Launch pop-up "Holiday Izakaya" concepts in Italian/Japanese restaurants (low-holiday genres).

Target: Increase weekday per-customer spend by 15%, capture 30% of holiday demand from top competitors.

5. **Double Visitors in 6 months**

Phase 3: Break 500 Visitors/Day (Month 5-6)

Problem: Requires +250 daily visitors above natural growth.

1. LSTM Model Enhancement:
   - Add features: holiday_flag, local_events, weather from public APIs.
   - Output: Promo-trigger system (IF forecast_dip > 15% → SMS 15% off coupon).
2. Network Effects:
   - Launch referral program: "Bring 4 friends → 5th dines free" (math: 20% group size increase = 50+ visitors/day).
   - Algorithm: Track virality via K-factor = (referrals × conversion_rate).

Target: Achieve K-factor ≥ 0.8 (exponential growth).

💡 Key Insight: Doubling requires converting volatility into opportunity. The 78% weekly drops aren't just risks—they're discounting windows to acquire price-sensitive customers.

6. If these restaurants were in your city, what other data would you join to get more insights and increase visitors?

City: Mexico City

To unlock deeper insights for restaurants in Mexico City (CDMX), I'd prioritize integrating these local data sources, ordered by impact potential:

1. CDMX-Specific Holiday & Event Calendar

- Japan's holidays ≠ Mexico's (e.g., Día de Muertos, Grito de Independencia).
- Join with: `date` field to analyze visitor spikes during *local* festivities like:
  - Zócalo events
  - CDMX Restaurant Week
  - Lucha Libre nights

2. Proximity to Metro/Mobility Hubs

- Traffic congestion alters dining behavior. Integrate:
  - Metro/Metrobús station locations (e.g., "X restaurant is 300m from Bellas Artes station")
  - EcoBici docking station usage data
  - Didi pickup density by hour

    Join with: `restaurant_id` + `timestamp` to optimize promotions around commute peaks.

6. If these restaurants were in your city, what other data would you join to get more insights and increase visitors?

3. Safety Perception Scores

- Nighttime traffic drops in high-crime colonias. Add:
  - INEGI safety surveys by neighborhood
  - User-generated safety tags (e.g., "safe for solo dining" on Google Maps)
  - Police incident reports

  Join with: `area_name` + `visit_hour` to target security-enhanced promotions.

4. Weather & Pollution Sensitivity

- Rainy seasons and "contingencia ambiental" days crush outdoor dining. Track:
  - Hourly rainfall/UVI index (SMN)
  - IMECA air quality alerts
  - Temperature/humidity swings

  Join with: `visit_date` to trigger "rainy day discounts" or indoor seating promos.

6. If these restaurants were in your city, what other data would you join to get more insights and increase visitors?

5. Cultural/Tourism Micro-Trends

- CDMX's tourism surge (Condesa/Roma) demands:
  - Airbnb density by neighborhood
  - Walking tour routes (e.g., "Coyoacán Frida Kahlo trails")
  - Concert/event schedules (Foro Sol, Palacio de los Deportes)

    Join with: `restaurant_location` to capture tourist vs. local patterns.

7. DiDi Rides App Download Channels

Acquisition & Quality Framework

Channel Type	Examples	Quality Metrics	Assessment Method
📍 Paid	• Google/FB Ads • OEM Pre-installs (Xiaomi) • KOL Campaigns	• CPI • D7 Retention • LTV Efficiency	A/B Testing Cohort Analysis
🚀 Owned (Best)	• DiDi Food Cross-Promo • Driver QR Codes • Email/SMS	• Cross-Install Rate • Blended LTV • Scan Conversion	SDK Tracking Attribution Platforms
📣 Earned	• App Store Optimization • User Referrals • PR Coverage	• Organic Share • K-factor • Brand Searches	Promo Codes Analytics Dashboards

7. DiDi Rides App Download Channels

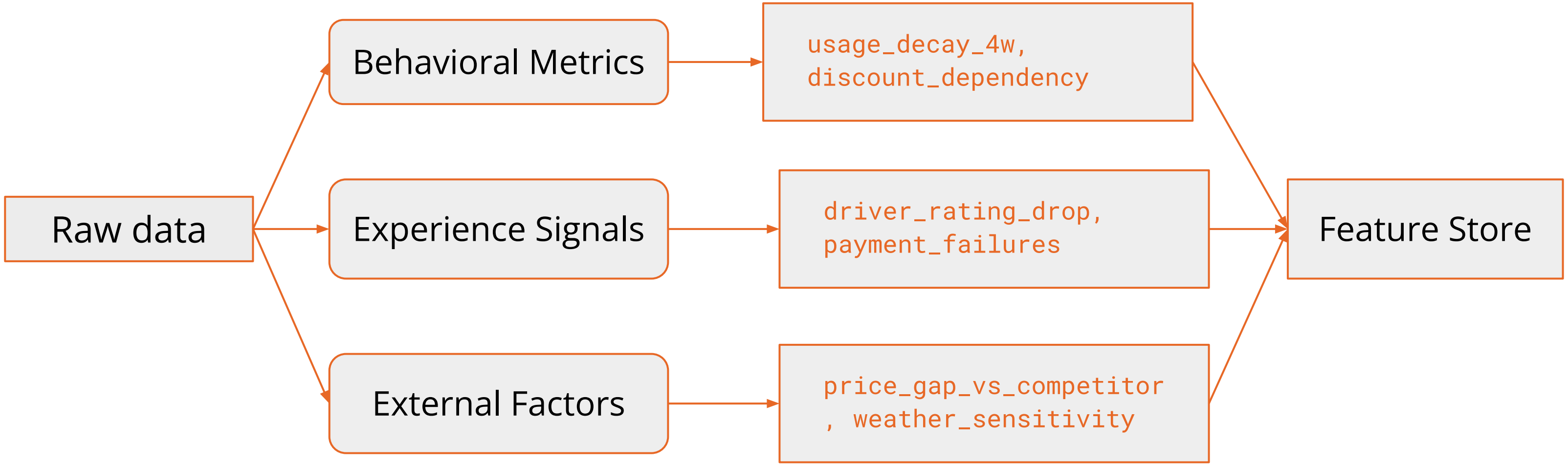
Cost Optimization & Channel Prioritization

Channel	Cost Metric	Efficiency Formula	ROAS Target
OEM Pre-installs	$0.30/install	LTV / CPI	230%
DiDi Food Cross-Promo	Near-zero CPI	Incremental Revenue / Cost	400%+
Driver QR Codes	$1.50/referral	Ride Completion Rate × LTV	180%
TikTok Ads	$2.80 CPI (Brazil)	6-mo Revenue / Ad Spend	150%

8. Churn Prediction Model for DiDi Rides APP

Transforming Raw Data into Churn Signals

Data pipeline diagram



8. Churn Prediction Model for DiDi Rides APP

Key Table

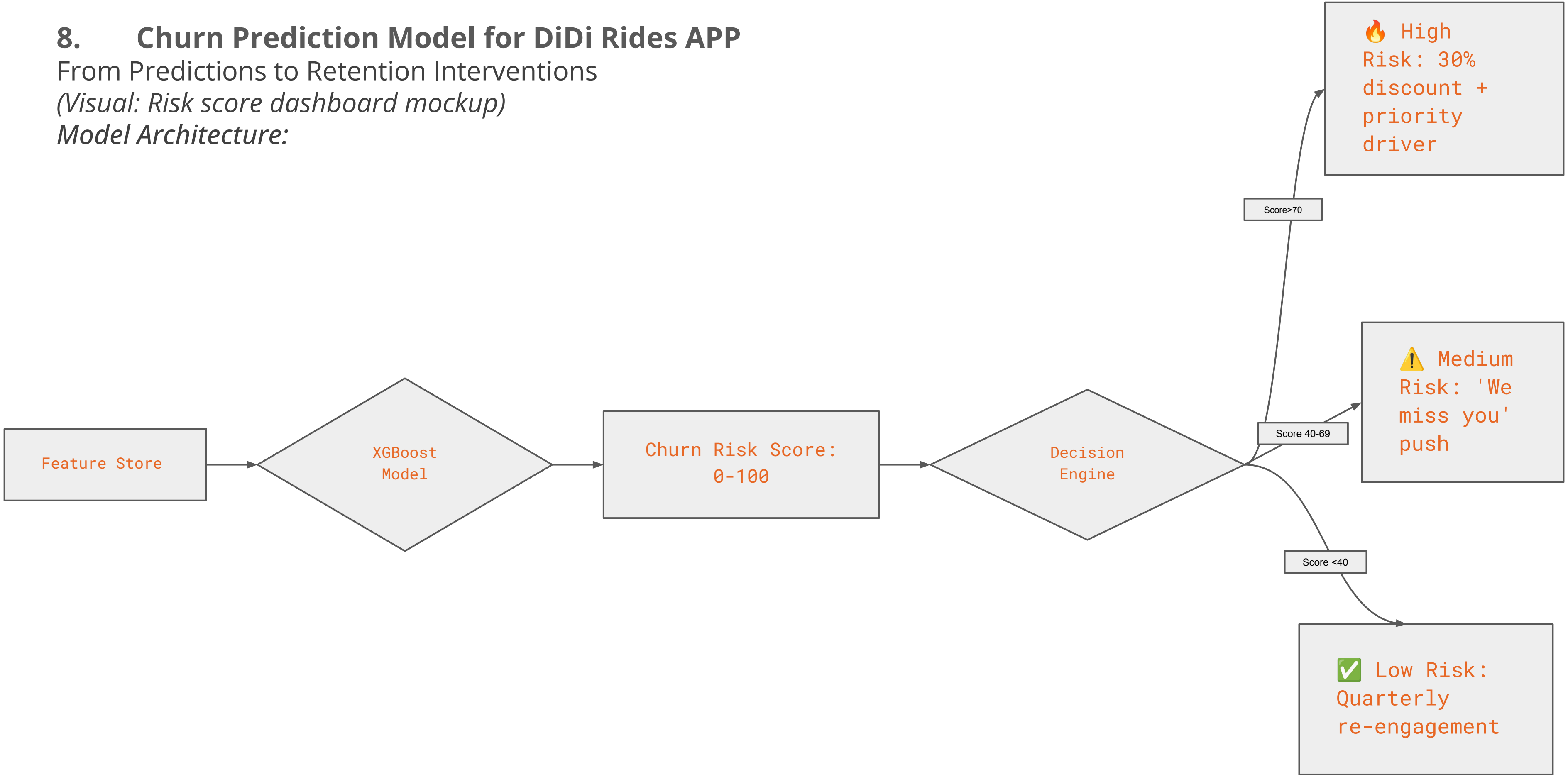
Feature Type	Example Metrics	Churn Risk Impact
Behavioral	`usage_decay_4w`, `weekend_usage_ratio`	45% of predictive power
Experience	`driver_rating_avg`, `support_tickets`	30%
External	`competitor_app_installed`, `local_events`	25%

8. Churn Prediction Model for DiDi Rides APP

From Predictions to Retention Interventions

(Visual: Risk score dashboard mockup)

Model Architecture:



8. Churn Prediction Model for DiDi Rides APP

Intervention Effectiveness:

Risk Tier	Expected Retention Lift	Cost per Saved User
High (70-100)	42%	$3.20
Medium (40-69)	28%	$1.80
Low (<40)	9%	$0.50

Why XGBoost:

- Handles 100+ feature types (numeric/categorical)
- Computes feature importance:
  `usage_decay_4w`: ⭐⭐⭐⭐⭐ (34% impact)
  `price_gap_vs_uber`: ⭐⭐⭐⭐ (22% impact)