



Abschlussreport

Projekt: *Favorita – Grocery Sales Forecasting*

Autorin: *Julia Felgentreu, Data Analyst*

Zeitraum: *Q1 2014 (Train: 2013–Februar, Forecast: März 2014)*



Zielsetzung

Ziel dieses Projekts war es, die täglichen Verkaufszahlen der ecuadorianischen Supermarktkette *Corporación Favorita* für die Region **Guayas** vorherzusagen. Im Fokus stand die Entwicklung eines **sauberen, reproduzierbaren und nachvollziehbaren Forecast-Systems**, das Demand Plannern eine verlässliche Grundlage für operative Entscheidungen bietet – insbesondere im Hinblick auf Personalplanung, Warenverfügbarkeit und Promotions.



Vorgehensweise

1. Datenaufbereitung

Die Rohdaten aus dem Kaggle-Datensatz wurden auf **Stores der Region Guayas** gefiltert und zu einem konsistenten Zeitreihenformat transformiert.

Dabei wurden:

- negative Verkaufswerte auf 0 gesetzt,
- fehlende Datumswerte über einen vollständigen Kalender pro Store×Item ergänzt,
- externe Tabellen (Oil, Transactions, Holidays, Items, Stores) über das Datum gemerged,
- der Ölpreis linear interpoliert, um Ausreißer und Gaps zu vermeiden.

So entstand ein sauberer, lückenfreier Datensatz mit hoher Datenqualität — die Basis für Feature Engineering und Modellierung.

2. Explorative Datenanalyse (EDA)

In der **explorativen Phase** wurden saisonale und verhaltensbezogene Muster sichtbar:

- Deutlich **höhere Verkaufszahlen an Wochenenden**, vor allem samstags.
- **Promotions** führen zu kurzfristigen Peaks und sind ein entscheidender Treiber der Nachfrage.
- Der **Ölpreis** zeigt nur eine geringe Korrelation ($p \approx -0.64$) zum Absatz – relevant als globaler Trend, aber nicht als direkter Einflussfaktor.
- Ein klarer **Pareto-Effekt**: Ein kleiner Anteil der Items generiert den Großteil der Umsätze.

Diese Erkenntnisse bildeten die Grundlage für die spätere Feature-Auswahl und Validierungsstrategie.

3. Feature Engineering

Um die zeitlichen und externen Dynamiken abzubilden, wurden folgende Feature-Gruppen entwickelt:

- **Lag-Features**: 1, 7, 14 und 30 Tage Rückblick zur Modellierung kurzfristiger und zyklischer Muster.
- **Rolling Features**: Gleitende Mittelwerte und Standardabweichungen (z. B. `roll_mean_7`) zur Glättung von Peaks.
- **Datumskomponenten**: Wochentag, Monat, Wochenend-Indikator zur Erfassung saisonaler und wöchentlicher Effekte.
- **Externe Variablen**: Ölpreis, Transaktionen, Promotions, Feiertage.



Das Training erfolgte **zeitlich strikt getrennt**, um *Data Leakage* zu vermeiden:

- Train: 2013–Februar 2014
 - Validation: letzte Februar-Woche
 - Test: März 2014
-

4. Modellierung & Evaluation

Für die Modellierung wurde **XGBoost** gewählt – aufgrund seiner Robustheit gegenüber nicht-stationären Daten und seiner Fähigkeit, komplexe Muster mit begrenzten Hyperparametern zu erfassen.

Es wurden zwei Varianten getestet:

-  **Baseline-XGB** mit Standardparametern
-  **Tuned-XGB** mit Early Stopping, regulierten Lernraten und Rolling-Origin-Validation

Zusätzlich wurde ein kleines **LSTM-Modell** als Vergleichsbaseline eingesetzt. Das XGBoost-Modell erzielte die besten Ergebnisse:

Modell	MAE	RMSE	Bias	sMAPE
XGB Baseline	3.39	7.93	−0.12	51.95
XGB Tuned	3.42	7.90	−0.10	53.06
LSTM	6.98	13.49	+0.85	87.14

Der minimale Bias zeigt, dass das Modell sehr gut kalibriert ist.

5. Forecast-Ergebnisse (März 2014)

Die Vorhersagen für März 2014 treffen den **Gesamttrend präzise**. Das Modell reagiert stabil auf reguläre Nachfragebewegungen, leichte Abweichungen entstehen hauptsächlich:

- an **Promotagen**,
- bei **lokalen Nachfragepeaks**,
- oder bei **unvorhergesehenen Feiertagseffekten**.

Diese Abweichungen sind typisch und zeigen, wo zukünftige Modellverbesserungen ansetzen können.

6. Modell-Interpretation & Stabilität

Zur Erklärbarkeit wurden **SHAP-Werte** herangezogen:

- Haupttreiber sind die **rollierenden Mittelwerte** und **Lag-Features** der letzten 7–14 Tage.
- Die **Transactions** wirken lokal erklärend, besonders bei kurzfristigen Nachfrageänderungen.
- Die Kalender-Features bestätigen das Wochenmuster: Freitag und Samstag zeigen höhere Absatzwerte.

Ein **Drift-Check (PSI)** zeigte:

- leichte Verschiebungen beim **Ölpreis** (globaler Trend),
- moderate Volatilität bei **Transactions**.

Beides gilt als erwartungskonform und bestätigt die **Stabilität des Modells über den Zeitraum hinweg**.

Zentrale Erkenntnisse

- **Nicht-Stationarität ist in Retail-Zeitreihen normal** – Modelle wie XGBoost können diese Schwankungen zuverlässig abbilden.
 - **Promotions und Feiertage** verursachen den größten Forecast-Error → Potenzial für gezieltes Event-Modeling.
 - **Perishables (frische Produkte)** zeigen stärkere Volatilität → erfordern engmaschigere Kontrolle im Forecast.
 - **Feature Stability** und moderate Regularisierung sind entscheidend, um Overfitting zu vermeiden.
-

Handlungsempfehlungen

Kurzfristig (technisch)

- **Promo-Handling verbessern:** Promotions und Feiertage als eigenständige Ereignis-Features modellieren (binary flags oder embeddings).
- **Item-Clustering:** Gruppenähnlicher Produkte (z. B. nach Family oder Shelf-Life) für hierarchische Forecasts nutzen.
- **Rolling Retraining:** Wöchentliche Retrainings zur Drift-Anpassung einführen.

Mittelfristig (strategisch)

- **Forecasts operativ nutzbar machen:** Integration in ein Dashboard (z. B. Streamlit oder Power BI).
- **Perishables separat überwachen:** Fokus auf Reduktion von Food Waste durch präzisere Kurzfristprognosen.
- **Feedback-Loop etablieren:** Vergleiche zwischen Prognose und Real Sales automatisiert rückführen, um das Modell laufend zu verbessern.

Ausblick

Das Projekt liefert eine stabile Grundlage für zukünftige Forecast-Anwendungen. Eine **interaktive Streamlit-App** wurde vorbereitet, um die Vorhersagen visuell zu explorieren und den Forecast-Prozess für Planer greifbar zu machen. Langfristig kann das System zu einem skalierbaren, operativen **Retail-Forecasting-Tool** ausgebaut werden – mit Fokus auf Echtzeit-Daten und adaptiven Modellen.

Fazit

Dieses Projekt zeigt, wie **präzise Demand Forecasts** aus gut vorbereiteten Daten entstehen können. Es kombiniert analytische Strenge mit praktischer Anwendbarkeit – und beweist, dass auch mit überschaubaren Modellen wie XGBoost robuste, interpretierbare und geschäftsrelevante Ergebnisse erzielt werden können.

„Data is the foundation — but understanding demand is what creates value.“ 