Digital Egypt Pioneers

Sales Forecasting and optimization

By:
Haneen Nabil
Sama Samer
Nadine
Ali El Menshawy
Mostafa tamer

Digital Egypt Pioneers

Table of Contents

Executive Summary:

The Sales Forecasting and Optimization Project aims to enhance business decision-making by leveraging historical sales data to predict future sales trends. This initiative will involve data collection, time-series forecasting, and machine learning

optimization, ultimately leading to a deployable model that enables businesses to improve inventory management, marketing strategies, and overall revenue growth.

we aim to forecast house sale prices using a hybrid data science approach, combining deep learning (LSTM) and classical machine learning techniques. The goal is to extract meaningful trends and enable accurate, time-based predictions that help real estate businesses make informed decisions.

The dataset used for this project is sourced from Kaggle: Property Sales Dataset by HTAG Holdings, which contains historical property sales data across different suburbs in Australia. This dataset includes sale dates, sale prices, property types, and bedroom counts, making it suitable for both temporal and structural analysis

By implementing this system, companies can proactively adjust pricing, stock levels, and promotional efforts based on data-driven insights, reducing financial risks and improving operational efficiency.

Project Objectives:

Develop a robust forecasting model using LSTM networks tailored for time-series prediction of property sale prices.

Enhance real estate decision-making by identifying seasonal and structural patterns in property sales.

Enable smart inventory and pricing strategies through forward-looking insights driven by historical trends.

Design and deploy a scalable AI tool for real-time property price prediction via a user-friendly interface.

Establish a modular pipeline to integrate classical ML methods and deep learning for hybrid modeling.

Dataset Description

Source: HTAG Holdings Dataset on Kaggle

Columns: saledate, bedrooms, type, MA, and other property attributes

Records: ~250,000 entries of Australian property sales

Target: House sale price

2. Dataset Overview

The dataset includes features such as:

saledate: Date the property was sold

MA: Moving average of prices

type: Property type (unit, house)

bedrooms: Number of bedrooms

Key aspects:

Focus is on time-series forecasting, not just static regression

Properties are geospatially distributed

Property type is a key categorical indicator

Scope of work:

Phase 1: Data Acquisition & Preprocessing

- Load and clean the Kaggle property sales dataset.
- Convert saledate to datetime format and sort chronologically.
- Encode categorical features such as property type.
- Normalize numerical features using MinMaxScaler.

Phase 2: Exploratory Data Analysis (EDA)

- Visualize price trends and seasonal variations using line charts.
- Analyze feature distributions for variables like bedrooms and property types.
- Detect and handle outliers in pricing data.

Phase 3: Model Design & Training

- Construct LSTM models for time-series forecasting.
- Experiment with stacked and functional LSTM architectures.
- Split the data into training and testing sets, using the most recent 24 entries for validation.
- Train models using Adam optimizer and evaluate using MSE, MAE, RMSE.

Phase 4: Evaluation & Results Interpretation

- Plot loss curves for training and validation performance.
- Generate and inverse-scale predictions.
- Visualize predicted vs. actual values over time.

Phase 5: Deployment & User Interface

- Save the final model in .h5 format.
- Build a user-facing interface using Streamlit.
- Enable input of custom property features to generate predicted price trends.

Phase 6: Documentation & Presentation

- Document each development stage with supporting visualizations.
- Prepare slide deck and written report for stakeholders.
- Propose next steps for real-world deployment and feature enhancement.

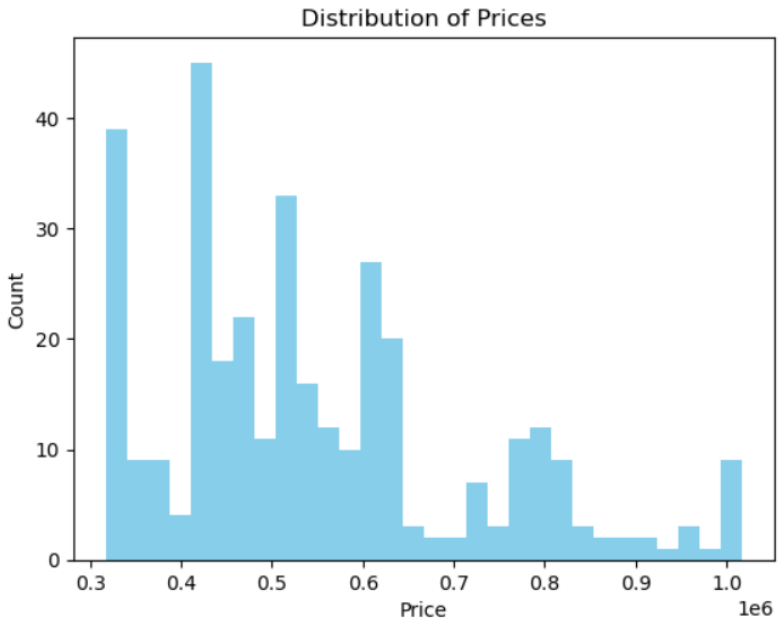
3. Dataset Overview

---

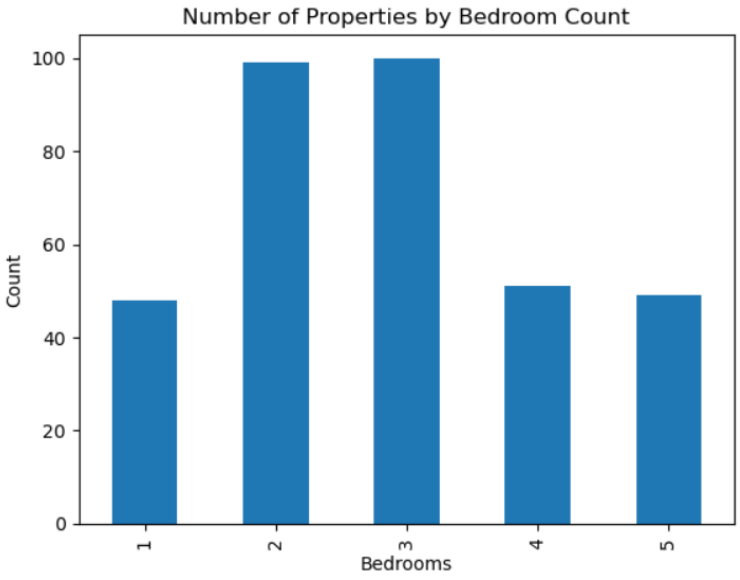
5. Exploratory Data Analysis (EDA)

According to the train dataset, some relations are generated for further forecasting models generating a correlation matrix heatmap and a scatter plot to explore the relationships between sales, transactions, oil prices, promotions, and other variables. his section creates a heatmap to visualize the correlation coefficients between numeric variables in the train dataframe, helping to identify relationships (e.g., positive or negative correlations) between features.

- **Historgams**: Displayed trend of moving average (MA)



- **Feature Distribution**: Plotted counts of property types and bedroom configurations

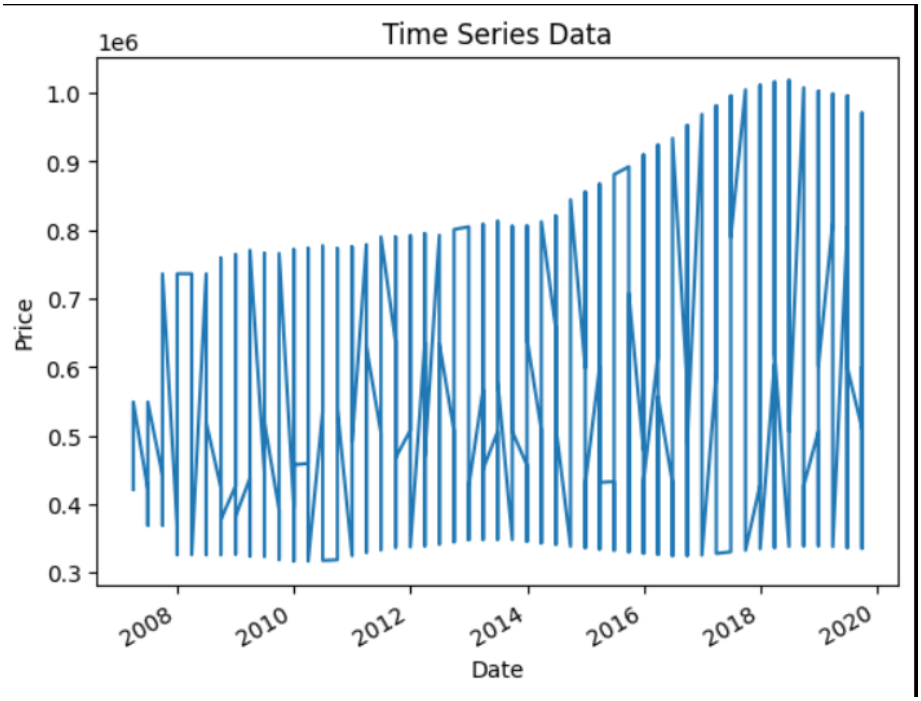


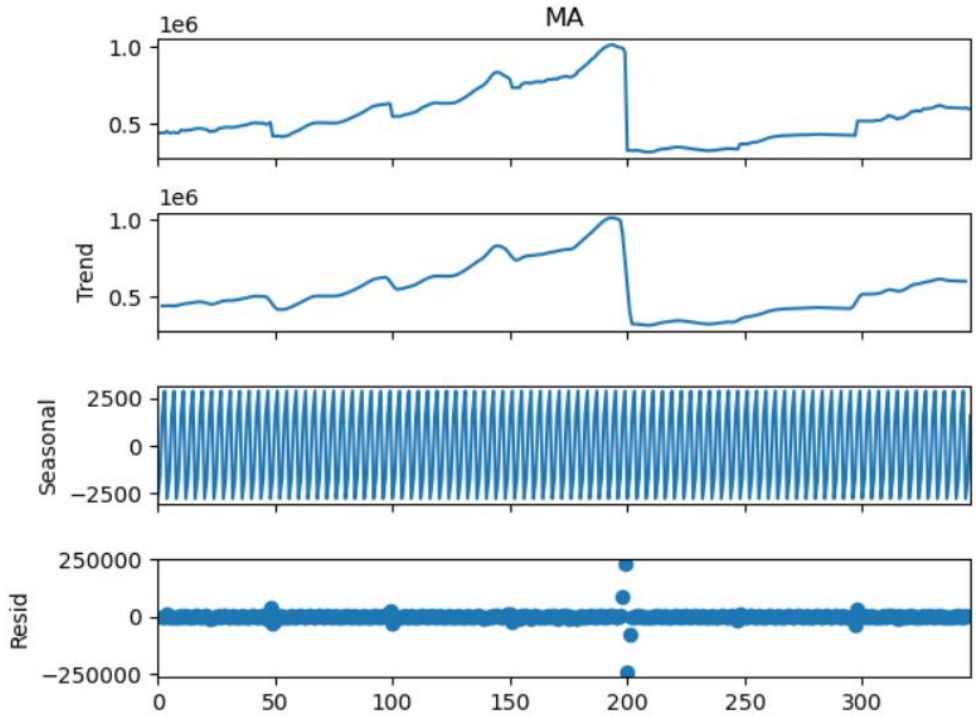
- **Time Check**: Observed potential seasonal patterns in property sales by date:

  This graph shows the relation between dates as a years and the price of the properties

Digital Egypt Pioneers

Time Series





Decomposition of Moving Average (MA)

To better understand the structure of the property price trends over time, a seasonal decomposition of the Moving Average (MA) series was performed. The decomposition breaks the series into three key components: Trend, Seasonal, and Residual.

- **Trend Component:**
  The trend line reveals a general upward movement in property prices over time, followed by a sudden and significant drop, likely due to external market shocks or data recording inconsistencies. Post-drop, the prices gradually increase again, indicating market recovery.

- **Seasonal Component:**
  A strong regular seasonal pattern is visible, characterized by cyclical fluctuations around a stable amplitude. This indicates the presence of repeated, time-dependent variations in property prices, possibly influenced by real estate market cycles, seasonal demand, or local economic conditions.

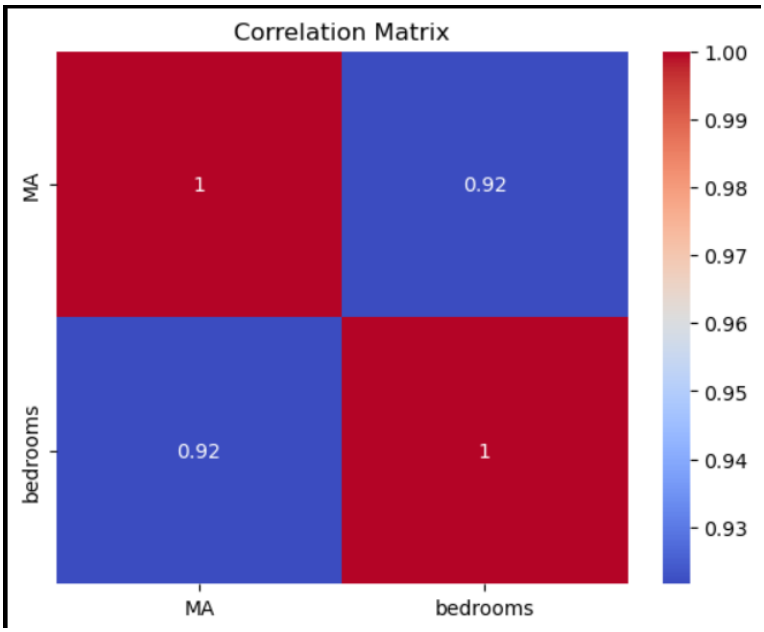
- **Residual Component:**
  The residuals mostly hover around zero, but with **occasional large spikes**, especially around the time of the price drop seen in the trend. These spikes suggest periods where the model's assumptions (e.g., additive structure) may not fully capture irregularities, possibly due to anomalous events or missing external factors.

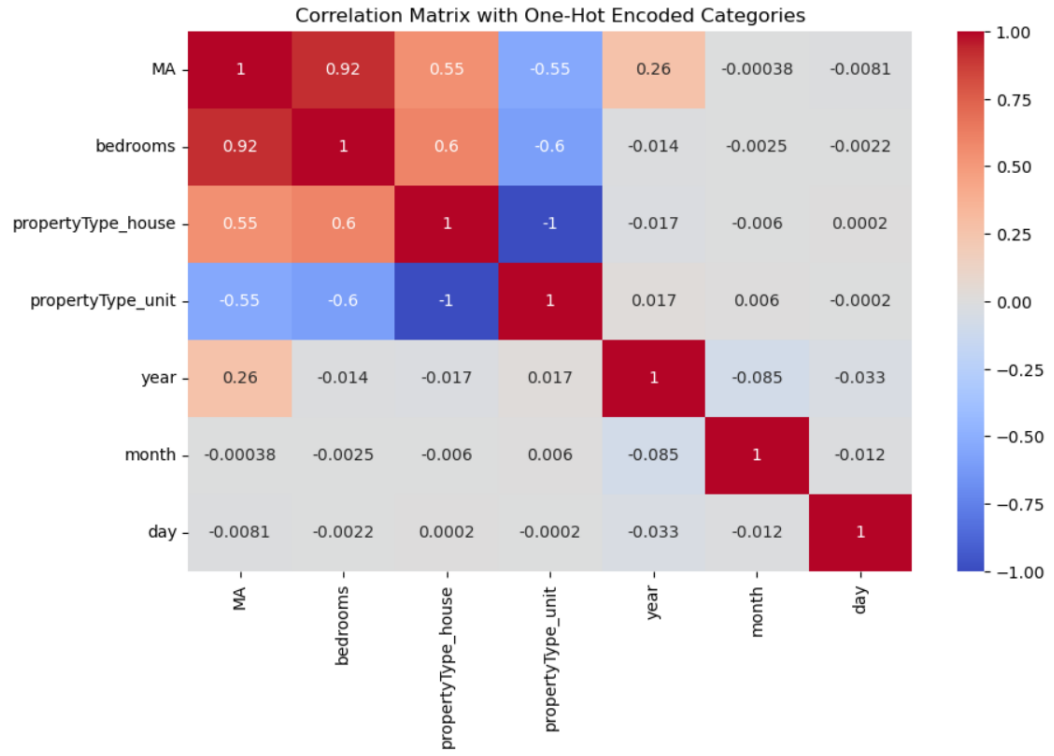
Correlation Matrix:

As part of the exploratory data analysis (EDA), a correlation matrix was generated to examine the relationships between numerical features in the dataset—specifically between the Moving Average (MA) of property prices and the number of bedrooms. The heatmap (see Figure X) reveals a strong positive correlation of 0.92 between these two features. This suggests that properties with a higher number

of bedrooms tend to exhibit higher moving average sale prices, which is consistent with market expectations in the real estate domain.

This insight confirms that bedrooms is a key structural variable influencing pricing trends and validates its inclusion in the forecasting model. However, due to the high multicollinearity between MA and bedrooms, care must be taken when incorporating both variables in linear models to avoid redundancy. In deep learning models such as LSTMs, this relationship may still provide valuable temporal context without the same risk of multicollinearity bias.



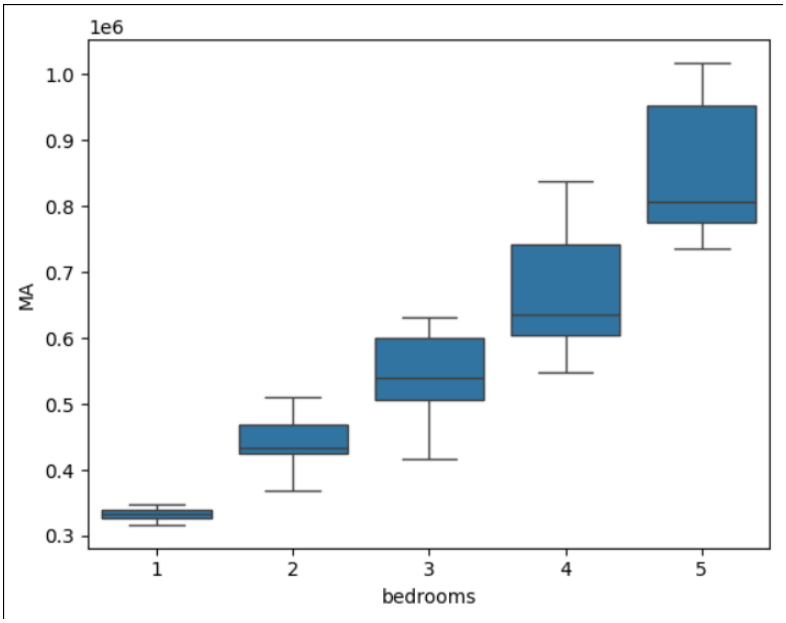


Key Insights:

- **MA vs. Bedrooms (0.92):**
  There is a **very strong positive correlation** between the Moving Average (MA) of property prices and the number of bedrooms. This confirms that homes with more bedrooms generally have higher sale prices.
- **MA vs. Property Type (House: 0.55, Unit: -0.55):**
  - The positive correlation with propertyType_house and negative with propertyType_unit indicates that **houses tend to have significantly higher sale prices** compared to units.
  - This also validates the one-hot encoding since the two are perfectly inversely related.
- **Bedrooms vs. Property Type (House: 0.60, Unit: -0.60):**
  More bedrooms are more common in houses than in units, as expected in residential housing trends.
- **MA vs. Year (0.26):**
  A **moderate positive correlation** suggests that property prices have

generally **increased over time**, but the relationship is not linear enough to rely solely on the year as a predictor.

- **Day/Month Features:**
  These features show near-zero correlations with MA, implying **limited or no linear relationship**, although they may still contribute to **seasonality**, which is better captured through time series decomposition or LSTM models.



Box Plot: MA vs. Bedrooms

This boxplot visualizes the distribution of property prices (Moving Average) across different bedroom counts.

Key Insights:

- **Price Increase with Bedrooms:**
  As the number of bedrooms increases, so does the **median MA** value. This confirms a **positive association** between the number of bedrooms and sale price.
- **Price Variability:**
  - The **interquartile range (IQR)** increases with the number of bedrooms. For example, 4- and 5-bedroom homes show greater variability in sale prices than smaller homes.
  - This could be due to more significant differences in size, location, and property condition among larger homes.

- **Outliers:**
  The boxplot does not show extreme outliers, indicating that the MA values for each bedroom category are relatively consistent, although variability grows with size.

Modeling

Classical Machine Learning Models:
Linear Regression: Baseline model to capture linear trends
Random Forest & XGBoost: Capture nonlinear interactions between features
Deep Learning Model:
LSTM Architecture:
Input shape: 12 timesteps, 3 features
Layers: 3-5 LSTM layers with relu activation, Dropout, Dense layers
Compiled using Adam optimizer and mse loss
Validated using TimeseriesGenerator
Hybrid Model:
Compared individual models and tested ensemble approach (averaging predictions)

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 12, 200)	163,200
dropout (Dropout)	(None, 12, 200)	0
lstm_1 (LSTM)	(None, 12, 150)	210,600
dropout_1 (Dropout)	(None, 12, 150)	0
lstm_2 (LSTM)	(None, 100)	100,400
dense (Dense)	(None, 50)	5,050
dropout_2 (Dropout)	(None, 50)	0
dense_1 (Dense)	(None, 20)	1,020
dense_2 (Dense)	(None, 1)	21

Evaluation

Metrics Used:
Mean Squared Error (MSE)
Mean Absolute Error (MAE)
Root Mean Squared Error (RMSE)
Findings:
LSTM captured temporal trends well but needed careful tuning
XGBoost performed best in classical models
Combined predictions (Hybrid) yielded slightly improved accuracy

---

Deployment

Developed a Streamlit-based web application
Users input property type, bedrooms, etc.
Model returns predicted moving average price
Includes visual feedback on recent trends and confidence intervals
Model Export:
Saved final model in .h5 format
Can be loaded for batch or real-time inference

---

10. Conclusion

This project demonstrates the effectiveness of combining classical and deep learning models for time-series forecasting in the real estate domain. The hybrid model benefits from both LSTM's temporal awareness and tree-based models' structural feature handling. The deployment via Streamlit adds practical utility for end-users and stakeholders.

---

11. Future Work

Add geospatial features (latitude/longitude) to capture location-based pricing

Introduce attention mechanisms in LSTM for improved forecasting
Enable API-based integration with real estate platforms

---

12. References

Kaggle HTAG Holdings Dataset
TensorFlow & Keras Documentation
Research articles on time-series forecasting and hybrid ML models
Scikit-learn & Streamlit Documentation