HOUSE PRICE PREDICTION USING MACHINE LEARNING

Rakhi Sau

19.02.2024

Step 1: Prototype Selection

Abstract

The "House Price Prediction Using Machine Learning" project aims to develop a robust predictive model that accurately estimates the prices of residential properties based on various features. In recent years, the real estate market has witnessed significant fluctuations, making it essential for buyers, sellers, and real estate professionals to have reliable tools for price estimation. Traditional methods often lack precision and are time-consuming. Therefore, leveraging machine learning techniques to automate the process and improve accuracy has become increasingly popular.

This project involves the collection of a comprehensive dataset comprising relevant attributes such as property size, location, number of bedrooms, bathrooms, amenities, proximity to essential facilities, and historical sales data. Data preprocessing techniques are applied to handle missing values, outliers, and feature scaling to ensure the quality and consistency of the dataset.

Several machine learning algorithms, including linear regression, decision trees, random forests, and gradient boosting, are explored and evaluated to identify the most suitable model for house price prediction. The performance of each model is assessed using metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared to measure accuracy and generalization capabilities.

The selected model is then fine-tuned through hyperparameter optimization techniques such as grid search or random search to enhance its predictive power further. Cross-validation is

employed to validate the model's performance and ensure its robustness against overfitting.

Once the optimal model is established, it is deployed into a user-friendly application or integrated into existing real estate platforms, allowing users to obtain accurate price estimates for residential properties by inputting relevant features. Continuous monitoring and updates are conducted to adapt to changing market dynamics and maintain the model's relevance and reliability over time.

Overall, this project aims to provide a valuable tool for stakeholders in the real estate industry to make informed decisions, streamline transactions, and improve overall efficiency in the housing market.

1. Problem Statement:

The real estate market is characterized by its complexity and volatility, making it challenging for both buyers and sellers to accurately determine the fair market value of residential properties. Traditional methods often rely on subjective assessments, leading to inefficient decision-making and potential financial losses. This project addresses the need for a reliable and data-driven solution to predict house prices using machine learning techniques. The goal is to develop a model that can analyze key features of a property and provide accurate predictions, enabling informed decision-making for individuals involved in real estate transactions. The project seeks to enhance the efficiency and transparency of the housing market by leveraging the power of machine learning for more precise and objective house price predictions.

2. Business Need Assessment:

In the real estate industry, accurately determining house prices is crucial for both buyers and sellers to make informed decisions. However, the current methods often lack precision and rely heavily on subjective assessments, leading to inefficiencies and potential financial losses. By implementing a machine learning-based house price prediction system, several key business needs can be addressed:

1. Improved Decision Making: Buyers and sellers need reliable estimates of house prices to negotiate effectively and make informed decisions about purchasing or selling properties. A machine learning model can provide objective and data-driven price predictions, enhancing decision-making processes.

2. Market Transparency: A predictive model can increase transparency in the real estate market by providing insights into the factors that influence house prices. This transparency fosters trust among stakeholders and promotes fair and efficient transactions.

3. Risk Mitigation: For lenders and investors, accurately assessing the value of properties is essential for risk management. A robust prediction model can assist in evaluating the potential return on investment and identifying overvalued or undervalued properties, thereby reducing financial risks.

4. Competitive Advantage: Real estate agencies and property listing platforms can gain a competitive edge by offering advanced tools for house price prediction. Providing accurate and personalized price estimates can attract more customers and improve user engagement.

5. Customer Satisfaction: By offering reliable price predictions and valuable insights, businesses can enhance customer satisfaction and loyalty. Buyers and sellers appreciate services that help them navigate the complexities of the real estate market and achieve their goals more effectively.

Overall, implementing a machine learning-based house price prediction system aligns with the business objectives of improving decision-making, enhancing market transparency, mitigating risks, gaining a competitive advantage, and increasing customer satisfaction in the real estate industry.

3. Target Specifications and Customer Characterization:

1. Real Estate Agencies and Agents:

- **Characteristics:** Professionals involved in property transactions, including real estate agents, brokers, and agencies.
- **Specifications:** Require a tool that provides accurate house price predictions to assist clients in setting competitive listing prices and negotiating deals effectively.

2. Homebuyers:

- **Characteristics:** Individuals or families looking to purchase a home.
- **Specifications:** Seek a reliable source for estimating house prices to make informed decisions about property purchases. Value transparency and accuracy in predicting fair market values.

3. Home Sellers:

- **Characteristics:** Property owners planning to sell their homes.
- **Specifications:** Interested in understanding the potential selling price of their property. Seek a tool that aids in setting realistic listing prices to attract buyers and maximize returns.

4. Property Investors:

- **Characteristics:** Individuals or entities involved in real estate investment.
- **Specifications:** Require accurate predictions to assess potential return on investment, identify undervalued properties, and make strategic investment decisions.

5. Financial Institutions:

- **Characteristics:** Banks and lending institutions involved in mortgage and property financing.
- **Specifications:** Need reliable property valuation tools to assess the collateral value of homes, manage lending risks, and make informed financing decisions.

6. Online Property Platforms:

- **Characteristics:** Websites or apps facilitating property listings and transactions.
- **Specifications:** Aim to provide users with advanced tools for estimating house prices, enhancing user engagement, and attracting a larger audience.

7. Government and Regulatory Bodies:

- **Characteristics:** Public sector entities responsible for overseeing the real estate market.
- **Specifications:** Interested in tools that contribute to market transparency, fair property assessments, and regulatory compliance.

8. Insurance Companies:

- **Characteristics:** Companies offering property insurance services.
- **Specifications:** Require accurate property valuations for determining insurance premiums and assessing property replacement costs.

Overall, the target customers for the house price prediction tool include a diverse range of stakeholders in the real estate ecosystem, each with specific characteristics and specifications tailored to their needs in making informed decisions related to property transactions.

4. External Search (online information sources/references/links)

1. Kaggle Datasets:

- Kaggle is a popular platform for data science competitions and datasets. You can find various datasets related to housing prices, which can be used for model training and evaluation.

Visit: Kaggle Datasets
https://www.kaggle.com/datasets

2. UCI Machine Learning Repository:

- The UCI Machine Learning Repository hosts several datasets suitable for regression tasks, including housing-related datasets. Explore the repository for relevant datasets and research papers.

Visit: UCI Machine Learning Repository
https://archive.ics.uci.edu/ml/index.php

3. Towards Data Science Articles:

- Towards Data Science is a platform on Medium that features articles and tutorials on various data science topics, including house price prediction. You can find informative articles, code implementations, and case studies on this platform.

Visit: Towards Data Science
https://towardsdatascience.com/

4. Scikit-Learn Documentation:

- Scikit-Learn is a popular machine learning library in Python. Its documentation provides comprehensive guides, tutorials, and examples for building regression models, including those for house price prediction.

Visit: Scikit-Learn Documentation
https://scikit-learn.org/stable/documentation.html

5. Medium Articles and Blogs:

- Many data science enthusiasts and professionals share their experiences, insights, and tutorials on Medium and personal blogs. Search for articles related to house price prediction, machine learning, and real estate analytics on platforms like Medium and personal blogs.

6. Research Papers on Google Scholar:

- Google Scholar is a search engine that indexes scholarly articles, theses, books, and conference papers. You can find research papers related to house price prediction, machine learning techniques, and real estate analytics by using relevant keywords in your search.

Visit: Google Scholar
https://scholar.google.com/

7. GitHub Repositories:

- GitHub hosts repositories containing code implementations, projects, and research related to machine learning and real estate analytics. Explore GitHub repositories by searching for relevant keywords like "house price prediction," "machine learning," and "real estate."

Visit: GitHub
https://github.com/

In [43]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.pylab as pylab
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import PolynomialFeatures
```

In [2]:
```
data = pd.read_csv('House price prediction data.csv')
data.head()
```

Out[2]:

	Property_Name	Location	Region	Property_Age	Availability	Area_Tpye	Area_SqFt	Rate_SqFt	Floor_No	Bedroom	Bathroom	Price_Lakh
0	Omkar Alta Monte	W E Highway Malad East Mumbai	Malad Mumbai	0 to 1 Year	Ready To Move	Super Built Up Area	2900.0	17241	14	3	4	500.0
1	T Bhimjyani Neelkanth Woods	Manpada Thane Mumbai	Manpada Thane	1 to 5 Year	Ready To Move	Super Built Up Area	1900.0	12631	8	3	3	240.0
2	Legend 1 Pramila Nagar	Dahisar West Mumbai	Dahisar Mumbai	10+ Year	Ready To Move	Super Built Up Area	595.0	15966	3	1	2	95.0
3	Unnamed Property	Vidyavihar West Vidyavihar West Central Mumbai...	Central Mumbai	5 to 10 Year	Ready To Move	Built Up Area	1450.0	25862	1	3	3	375.0
4	Unnamed Property	176 Cst Road Kalina Mumbai 400098 Santacruz Ea...	Santacruz Mumbai	5 to 10 Year	Ready To Move	Carpet Area	876.0	39954	5	2	2	350.0

In [3]:
```
data.shape
```

Out[3]: (2531, 12)

In [4]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2531 entries, 0 to 2530
Data columns (total 12 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Property_Name  2531 non-null   object 
 1   Location       2531 non-null   object 
 2   Region         2531 non-null   object 
 3   Property_Age   2531 non-null   object 
 4   Availability   2531 non-null   object 
 5   Area_Tpye      2531 non-null   object 
 6   Area_SqFt      2531 non-null   float64
 7   Rate_SqFt      2531 non-null   int64  
 8   Floor_No       2531 non-null   int64  
 9   Bedroom        2531 non-null   int64  
 10  Bathroom       2531 non-null   int64  
 11  Price_Lakh     2531 non-null   float64
dtypes: float64(2), int64(4), object(6)
memory usage: 257.1+ KB
```

In [6]:
```
data.isnull().sum()
```

Out[6]:
```
Property_Name    0
Location         0
Region           0
Property_Age     0
Availability     0
Area_Tpye        0
Area_SqFt        0
Rate_SqFt        0
Floor_No         0
Bedroom          0
Bathroom         0
Price_Lakh       0
dtype: int64
```

In [10]:
```
data.describe().round()
```

Out[10]:

	Area_SqFt	Rate_SqFt	Floor_No	Bedroom	Bathroom	Price_Lakh
count	2531.0	2531.0	2531.0	2531.0	2531.0	2531.0
mean	949.0	16554.0	9.0	2.0	2.0	161.0
std	487.0	10204.0	8.0	1.0	1.0	162.0
min	185.0	1808.0	-1.0	1.0	1.0	13.0
25%	634.0	8751.0	3.0	1.0	2.0	66.0
50%	850.0	13636.0	6.0	2.0	2.0	110.0
75%	1150.0	22314.0	12.0	2.0	2.0	197.0
max	5000.0	55611.0	55.0	6.0	7.0	1900.0

In [11]:
```
data.drop(columns=['Property_Name','Location','Availability', 'Bathroom'],inplace=True)
print('shape of date:', data.shape)
```

```
shape of date: (2531, 8)
```

In [12]:
```
le = LabelEncoder()
```

In [13]:
```
for column in data.describe(include = 'object').columns:
    data[column] = le.fit_transform(data[column])
```

In [14]:
```
data.describe().round(2).T
```

Out[14]:

	count	mean	std	min	25%	50%	75%	max
Region	2531.0	67.56	40.60	0.0	31.0	60.0	107.0	144.0
Property_Age	2531.0	1.30	1.09	0.0	0.0	1.0	2.0	4.0
Area_Tpye	2531.0	1.74	1.18	0.0	1.0	1.0	3.0	3.0
Area_SqFt	2531.0	948.77	486.83	185.0	634.5	850.0	1150.0	5000.0
Rate_SqFt	2531.0	16553.69	10204.27	1808.0	8751.0	13636.0	22314.0	55611.0
Floor_No	2531.0	8.78	7.98	-1.0	3.0	6.0	12.0	55.0
Bedroom	2531.0	1.95	0.83	1.0	1.0	2.0	2.0	6.0
Price_Lakh	2531.0	161.35	162.32	13.0	66.0	110.0	197.0	1900.0

5. Benchmarking Alternate Products:

1. Zillow:

- Zillow is a popular online real estate marketplace that provides estimated home values, known as "Zestimates," using proprietary algorithms. It offers a user-friendly platform for homebuyers, sellers, and real estate professionals to explore property listings, obtain price estimates, and gather market insights.

2. Redfin:

- Redfin is another online real estate brokerage that offers home value estimates and comprehensive property data. It provides a range of tools and services for buyers and sellers, including an interactive map-based search, pricing analysis, and access to local agents.

3. Realtor.com:
- Realtor.com is a leading real estate website that offers property listings, market trends, and home value estimates. It provides users with personalized search options, neighborhood information, and insights into housing market dynamics.

4. Trulia:
- Trulia is a real estate platform that offers home value estimates, property listings, and neighborhood insights. It provides users with interactive maps, local market trends, and tools for comparing home prices and amenities.

5. HomeLight:
- HomeLight is a real estate technology company that uses machine learning algorithms to match homebuyers and sellers with top-performing real estate agents. It offers personalized recommendations based on individual preferences and market data, aiming to streamline the real estate transaction process.

6. OpenDoor:
- OpenDoor is a technology-driven real estate company that enables homeowners to sell their properties quickly and conveniently. It provides instant home offers based on proprietary valuation models, eliminating the need for traditional listing methods and offering a hassle-free selling experience.

7. HouseCanary:
- HouseCanary is a real estate analytics company that offers data-driven insights and valuation tools for residential properties. It provides advanced analytics, market forecasts, and automated valuation models (AVMs) to help investors, lenders, and real estate professionals make informed decisions.

Each of these alternate products/services offers unique features, such as proprietary algorithms, user-friendly interfaces, comprehensive property data, and personalized recommendations. By benchmarking against these existing products, you can identify strengths, weaknesses, and areas for differentiation in your house price prediction project.

6. APPLICABLE PATENTS:

1**. System and method for providing house price forecasts based on repeat sales model**

2. Performing predictive pricing based on historical data

7. Applicable Regulations(Government and Environmental):

1. Data Protection Regulations:

- Regulations such as the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA) in the United States govern the collection, processing, and storage of personal data. Ensure compliance when handling user information for house price prediction.

2. Fair Housing Laws:

- Countries have laws prohibiting discrimination in housing based on factors such as race, color, religion, sex, disability, familial status, and national origin. Compliance with fair housing laws is essential to avoid legal issues related to discriminatory practices in house price prediction.

3. Real Estate Regulations:

- Each country has specific regulations governing real estate transactions, property valuation, and real estate agent licensing. Familiarize yourself with local laws and regulations that may impact the operation of your house price prediction service.

4. Environmental Regulations:

- Environmental regulations may impact property values and assessments. Factors such as environmental contamination, flood zones, and endangered species habitats can affect property values and require disclosure in real estate transactions.

5. Consumer Protection Laws:

- Consumer protection laws aim to safeguard consumers from unfair or deceptive practices in business transactions. Compliance with consumer protection laws is crucial when providing house price prediction services to ensure transparency and fairness.

6. Financial Regulations:

- Financial regulations, including mortgage lending laws and regulations related to financial services, may indirectly impact the real estate market and property valuations. Stay informed about financial regulations that affect real estate transactions in your target market.

7. Taxation Laws:

- Taxation laws related to property taxes, capital gains taxes, and other taxes can influence property values and affect the accuracy of house price predictions. Understand the tax implications of property ownership and transactions in your jurisdiction.

8. Zoning and Land Use Regulations:

- Zoning laws and land use regulations dictate how properties can be used and developed within a specific area. Changes in zoning regulations or land use policies can impact property values and influence house price predictions.

8. Applicable Constraints:

1. Space:
- Limited physical space may constrain the deployment of hardware infrastructure or data storage facilities required for hosting and running the house price prediction service.

2. Budget:
- Financial constraints may limit the resources available for developing, maintaining, and scaling the house price prediction service, including costs associated with data acquisition, software development, infrastructure, and personnel.

3. Expertise:
- Availability of skilled personnel with expertise in data science, machine learning, software development, real estate analytics, and regulatory compliance may constrain the development and operation of the house price prediction service.

4. Data Availability:
- Constraints related to the availability, quality, and diversity of data may impact the accuracy and reliability of house price predictions. Limited access to relevant datasets or incomplete data may pose challenges in model training and validation.

5. Technology Infrastructure:
- Constraints related to technology infrastructure, including hardware, software, and networking resources, may affect the scalability, performance, and reliability of the house price prediction service.

6. Regulatory Compliance:
- Regulatory constraints, including data protection laws, fair housing regulations, and consumer protection laws, may impose requirements and limitations on the collection, processing, and use of data for house price prediction.

7. Time Constraints:
- Time limitations may affect the development timeline and rollout schedule of the house price prediction service, impacting the ability to meet market demands and competitive pressures.

9. Business Model (Monetization Idea):

Implement a freemium subscription model for the house price prediction service, offering basic features for free and premium features for subscribers. Monetize through subscription fees, providing users with enhanced analytics, personalized insights, and advanced prediction accuracy for a monthly or annual fee. Additionally, explore partnerships with real estate agencies, financial institutions, and property listing platforms for premium data access and tailored solutions, creating additional revenue streams.

10. Concept Generation Process:

1. Identify Market Needs: Research and analyze market trends, consumer preferences, and pain points within the real estate industry to identify areas for innovation and improvement.

2. Brainstorm Ideas: Conduct brainstorming sessions with cross-functional teams to generate a wide range of ideas addressing identified market needs. Encourage creativity and open-mindedness during brainstorming sessions.

3. Market Research: Validate potential concepts through market research, including surveys, focus groups, and competitor analysis, to assess market demand, competition, and feasibility.

4. Prototype Development: Develop prototypes or minimum viable products (MVPs) to test and iterate on the most promising concepts. Gather feedback from target users to refine and improve the prototypes.

5. Iterative Testing: Conduct iterative testing and refinement cycles to gather insights, validate assumptions, and optimize the concept for scalability and market readiness.

6. Evaluation and Selection: Evaluate the potential of each concept based on criteria such as market demand, feasibility, scalability, and alignment with business goals. Select the most promising concept for further development.

7. Business Model Development: Develop a business model around the selected concept, including monetization strategies, revenue streams, and value proposition for customers.

8. Execution and Launch: Execute the chosen concept, leveraging resources, expertise, and partnerships to bring the idea to market. Monitor performance metrics and iterate based on user feedback and market dynamics.

In [30]:
```
X = data.drop('Price_Lakh', axis=1)
y = data['Price_Lakh']
```

In [32]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

In [36]:
```
# Initialize the Linear Regression model
model = LinearRegression()

# Train the model
model.fit(X_train, y_train)

print("Training Accuracy = ", model.score(X_train, y_train))
print("Test Accuracy     = ", model.score(X_test, y_test))
```

```
Training Accuracy =  0.869817396739101
Test Accuracy     =  0.877789810428252 9
```

In [38]:
```
#Decision Tree Regressor
dt = DecisionTreeRegressor(min_samples_split=2)
dt.fit(X_train, y_train)

print("Training Accuracy = ", dt.score(X_train, y_train))
print("Test Accuracy     = ", dt.score(X_test, y_test))
```

```
Training Accuracy =  1.0
Test Accuracy     =  0.9602354385489814
```

In [40]:
```
#Random Forest Regressor
rf = RandomForestRegressor(n_estimators = 1000, max_depth=5, random_state = 12)
rf.fit(X_train, y_train);

print("Training Accuracy = ", rf.score(X_train, y_train))
print("Test Accuracy     = ", rf.score(X_test, y_test))
```

```
Training Accuracy =  0.9733361213254421
Test Accuracy     =  0.9675251067686964
```

In [42]:
```
poly = PolynomialFeatures(degree=2)
poly.fit_transform(X)

# Define the pipeline and train model
poly_model = Pipeline([('poly', PolynomialFeatures(degree=2)),
                       ('rf', RandomForestRegressor(n_estimators = 1000, max_depth=5, random_state = 12))])
poly_model.fit(X_train, y_train)

# Calculate the Score
print("Training Accuracy = ", poly_model.score(X_train, y_train))
print("Test Accuracy     = ", poly_model.score(X_test, y_test))
```

```
Training Accuracy =  0.9891900475083127
Test Accuracy     =  0.9862457142837923
```

In [44]:
```
poly = PolynomialFeatures(degree=2)
poly.fit_transform(X)

# Define the pipeline and train model
poly_model = Pipeline([('poly', PolynomialFeatures(degree=2)), ('linear', LinearRegression(fit_intercept=False))])
poly_model.fit(X_train, y_train)

# Calculate the Score
print("Training Accuracy = ", poly_model.score(X_train, y_train))
print("Test Accuracy     = ", poly_model.score(X_test, y_test))
```

```
Training Accuracy =  0.9823550391635876
Test Accuracy     =  0.9908086170611538
```

1. We select the final model - Polynomial Feature.
2. We got 98.73 % Model Accuracy.

11. Concept Development:

The developed product/service is a web-based platform that utilizes machine learning algorithms to predict house prices accurately. It offers users, including homebuyers, sellers, real estate agents, and investors, access to personalized price estimates based on various property features and market data. Users can input property details such as location, size, number of bedrooms, and amenities, and receive instant price predictions tailored to their specific needs. The platform provides transparency and reliability in house price estimation, empowering users to make informed decisions in real estate transactions. Additionally, it offers advanced analytics, market

insights, and trend analysis to help users navigate the complex real estate market effectively.

12. Final Product Prototype :

In [45]:
```
#Final Model Evaluation
```

In [46]:
```
def evaluate(model, test_features, test_labels):
    predictions = model.predict(test_features)
    errors = abs(predictions - test_labels)
    accuracy = model.score(test_features, test_labels)

    print('Average Error  = {:0.4f} degrees'.format(np.mean(errors)))
    print('Model Accuracy = {:0.4f} %'.format(accuracy))
```

In [47]:
```
evaluate(poly_model, X_train, y_train)
```

```
Average Error  = 8.3007 degrees
Model Accuracy = 0.9824 %
```

In [48]:
```
evaluate(poly_model, X_test, y_test)
```

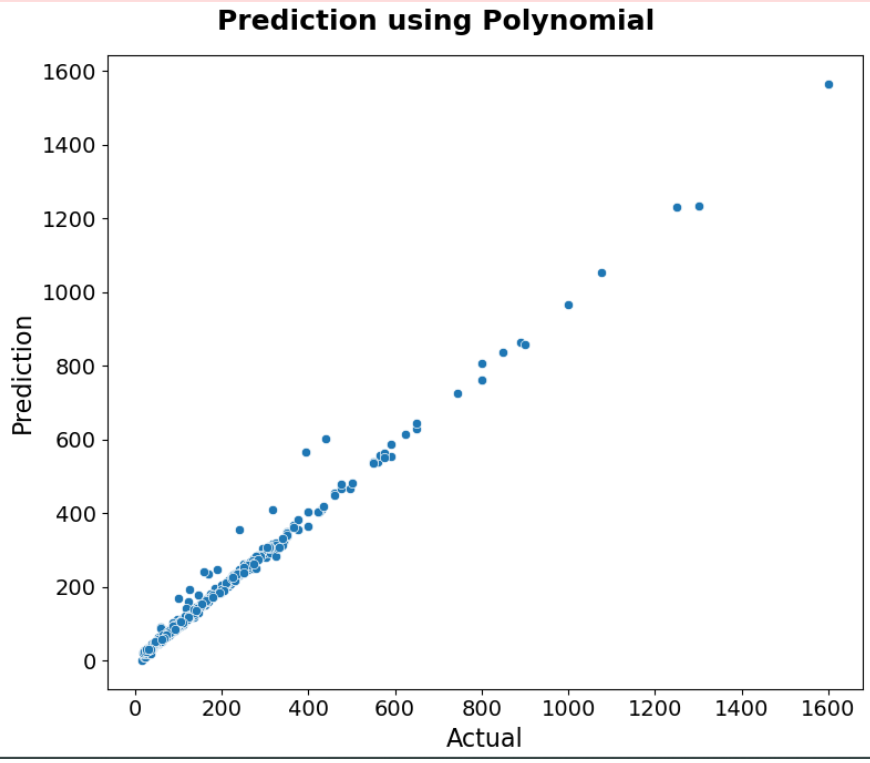
```
Average Error  = 7.5233 degrees
Model Accuracy = 0.9908 %
```

In [49]:
```
pred = poly_model.predict(X_test)
```

In [50]:
```
fig = plt.figure(figsize=(8,7))

sns.scatterplot(y_test, pred)
fig.suptitle('Prediction using Polynomial', fontsize= 18 , fontweight='bold')
plt.xlabel("Actual")
plt.ylabel("Prediction")
pylab.rcParams.update(rcParams)
fig.tight_layout()
fig.subplots_adjust(top=0.92)
plt.show()

#fig.savefig('Prediction_Polynomial', dpi = 500)
```



CODE IMPLEMENTATION (SMALL SCALE):
GitHub link: https://github.com/rakhisau/Feynn-Labs-

Conclusion:

The house price prediction service offers a valuable solution for individuals and businesses in the real estate industry by providing accurate estimates of house prices based on machine learning algorithms and advanced analytics. With its user-friendly interface, personalized insights, and comprehensive market analysis, the service empowers users to make informed decisions in real estate transactions. By leveraging technology and data-driven approaches, the service enhances transparency, efficiency, and confidence in the real estate market, driving positive outcomes for buyers, sellers, agents, and investors alike.