# PREDICTING HOUSE PRICES USING MACHINE LEARNING



#### **INDUPAMA HERATH**

#### SARASIJ MAITRA

#### **RAFATU SALIS**

### **ERSIN SÜER**



#### THE ERDŐS INSTITUTE

Helping PhDs get and create jobs they love at every stage of their career.

# **OBJECTIVES**

- Predict house prices incorporating traditional and non-traditional features using machine learning
- Build a web application to predict the house price for user input feature values





#### Stakeholders

- Families looking to settle down in King County, WA
- Real estate agents trying to give estimates to housing prices
- Local government agencies

### Why King County?

Land area: 2126 sq. mi. Water area: 180.5 sq. mi.

**Population density:** 1066 people per square mile (very high).

**Cost of living index**: 111.7 (more than average, U.S. average is 100)

- Presence of many companies such as Boeing, Amazon, Google, Microsoft, etc.
- Scenic parks, Golden Garden Beach
- Historic places such as Pike Place Market, Space-needle

# DATA SET



## DATA CLEANING

- Removed unnecessary columns and rows with missing values.
- Add new columns

• Age, Log\_price

Age = 2024 - YEAR\_BUILT Log\_price = Log(PRICE)

 Cleaned data set has 4700 rows and 19 columns (5 categorical and 14 numerical)

## EXPLORATORY DATA ANALYSIS PRICE DISTRIBUTION

Minimum: 49000 Mean: 1505699 Median: 998975 Maximum: 7000000

>

						Corr	elatior	n Heat	map							1.00
PRICE -	1	0.77	0.3	0.44	0.62	0.13	0.033	-0.033	0.15	0.71	-0.14	0.26	0.24	-0.019		1.00
log_price -	0.77	1	0.44	0.53	0.73	0.12	0.075	-0.075	0.11	0.59	-0.15	0.45	0.45	-0.14		0.75
BEDS -	0.3	0.44	1	0.8	0.72	-0.017	0.014	-0.014	0.084	-0.059	0.037	0.052	0.037	-0.069		
BATHS -	0.44	0.53	0.8	1	0.81	0.038	0.14	-0.14	0.11	0.062	-0.057	0.13	0.12	-0.045	-	0.50
SQUARE FEET -	0.62	0.73	0.72	0.81	1	0.14	0.071	-0.071	0.16	0.12	-0.051	0.17	0.15	-0.12		
LOT SIZE -	0.13	0.12	-0.017	0.038	0.14	1	-0.019	0.019	0.14	0.09	0.011	0.031	0.033	-0.089		0.25
YEAR BUILT -	0.033	0.075	0.014	0.14	0.071	-0.019	1	-1	0.027	-0.044	0.046	0.043	0.042	-0.19		
Age -	-0.033	-0.075	-0.014	-0.14	-0.071	0.019	-1	1	-0.027	0.044	-0.046	-0.043	-0.042	0.19		0.00
DAYS ON MARKET -	0.15	0.11	0.084	0.11	0.16	0.14	0.027	-0.027	1	0.074	-0.042	-0.06	-0.065	0.067	-	-0.25
\$/SQUARE FEET -	0.71	0.59	-0.059	0.062	0.12	0.09	-0.044	0.044	0.074	1	-0.17	0.34	0.34	0.043		
num_school -	-0.14	-0.15	0.037	-0.057	-0.051	0.011	0.046	-0.046	-0.042	-0.17	1	-0.29	-0.26	-0.16	-	-0.50
school_rating_mean -	0.26	0.45	0.052	0.13	0.17	0.031	0.043	-0.043	-0.06	0.34	-0.29	1	0.97	-0.36		
Bayes_RatingSchool -	0.24	0.45	0.037	0.12	0.15	0.033	0.042	-0.042	-0.065	0.34	-0.26	0.97	1	-0.31	-	-0.75
crime_percentage -	-0.019	-0.14	-0.069	-0.045	-0.12	-0.089	-0.19	0.19	0.067	0.043	-0.16	-0.36	-0.31	1		
	PRICE -	log_price -	BEDS -	BATHS -	SQUARE FEET -	LOT SIZE -	YEAR BUILT -	Age -	DAYS ON MARKET -	\$/SQUARE FEET -	num_school -	school_rating_mean -	Bayes_RatingSchool -	crime_percentage -		-1.00

#### CORRELATION ANALYSIS OF HOUSING MARKET VARIABLES

#### Primary tool: *seaborn*



Effect of Location on Log\_Price colored by Property Type

# MORE PLOTS

Primary tool: seaborn

Boxplot on the right shows the distribution of *log\_price* for the categorical variable property type.



Property Type



## MORE PLOTS



## FEATURE ENGINEERING

• Delete outliers based on price.

• One hot encoding on PROPERTY TYPE column

#### **PROPERTY TYPE**

Single Family Residential	3030
Townhouse	430
Condo/Co-op	160
Multi-Family (2-4 Unit)	63
Multi-Family (5+ Unit)	45
Mobile/Manufactured Home	32

## FINAL DATA SET

Data	columns (total 16 c	olumns):
#	Column	Non-Null Count
0	BEDS	3759 non-null
1	BATHS	3759 non-null
2	SQUARE FEET	3759 non-null
3	LOT SIZE	3759 non-null
4	zipcode	3759 non-null
5	LATITUDE	3759 non-null
6	LONGITUDE	3759 non-null
7	Bayes_RatingSchool	3759 non-null
8	crime_percentage	3759 non-null
9	Age	3759 non-null
10	Single Family	3759 non-null
11	Townhouse	3759 non-null
12	Condo	3759 non-null
13	Multi_Family4	3759 non-null
14	Multi_Family5	3759 non-null
15	log_price	3759 non-null

### 15 features to predict House Price (in log scale)

## MODELING

1.The baseline model – Average of the House Prices (log scale)
2.K – Nearest Neighbors
3.Multiple Linear Regression
4.Decision Tree
5.Random Forest
6.XGBoost

Used 5 – fold cross validation on the training set and RMSEs were computed for the predicted price in log scale.

## MODEL TUNING

RMSE for Different Models on the Training Set

- Used a grid search-based approach to find the best set of parameters for KNN, Decision Tree, Random Forest, and XGBoost.
- Best model, XGBoost gives RMSEs,

   0.0954 on training set
   0.1003 on testing set



### IMPORTANT FEATURES OF XGBOOST



### HISTOGRAMS FOR PREDICTED VS TRUE LOG PRICE

Histogram for predicted log price and True log price on the testing set



## **RESIDUAL PLOT**



# WEB APPLICATION

• Using our final model, we built a simple web application on Streamlit that takes in user inputs (relevant to our model) and predicts the house price.

 The app is publicly available at https://erdos-datascience-may2024realestate-projectnevrbzjn2sh2zgsrc.streamlit.app/



- Visualizations using streamlit\_folium
- clearer view of the surrounding big cities, airports, etc., to make informed choices

Enter desired Property Type

Condo/Co-op				~
Property type chosen b	by user: Condo/Co	-op		
Choose your desired zipc	ode			
98023				~
98001				
98002				
98003				
98004				
98005				
98006				
98007				
00000		CHINE I		



Chosen Latitude and Longitude (47.2906325, -122.2162074)



# CONCLUSIONS

- Overall, all models did improve from the baseline model and did well on predicting the price.
- Contribution of non-traditional features are similar to some of the traditional features.

## FUTURE WORK

- Extend the study for other states.
- Incorporate more relevant features such as
  - whether the house has experienced flooding,
  - has mold issues,
  - the quality of construction materials,
  - the floor plan, and
  - whether fixtures and appliances have been recently updated.

## ACKNOWLEDGEMENTS

Our team is very grateful for all the guidance and advice provided by our instructor Steven Gubkin, our mentor Alec Traaseth and Alec Clott throughout the course of this project. We also deeply thank Roman Holowinsky and The Erdős Institute for providing us the platform and opportunity to work on this project.

Ő

### THE ERDŐS INSTITUTE

Helping PhDs get and create jobs they love at every stage of their career.