# PREDICTING HOUSE PRICES IN KING COUNTY, WASHINGTON, USA

Panagiotis Petsas

This presentation is part of the course “Applied Data Science Capstone” from “IBM Data Science Professional Certificate” on Coursera.

# Table of contents

# Introduction

**Background**

- King County is one of the most important counties in USA
- It has a population of more than 2.25m people
- It covers almost 6000 $km^2$, where 490 of them being water
- It contains many big cities like Seattle as well as many historical places
- If we are looking for a house in King County, on what factors should we pay attention?

# Introduction

**State of the problem**

- Houses are one of the most important type of tangible assets
- Many factors can affect their price (house features, neighborhood features, spatial and temporal factors)
- Can we quantify how much each characteristic affect the house price?

**Individuals interested in this problem**

- Real estate agents, who want to determine the house prices for sell
- Individuals who want to buy or sell their house in the right price

# Data acquisition and manipulation

**Data sources**

- King County house records from Kaggle

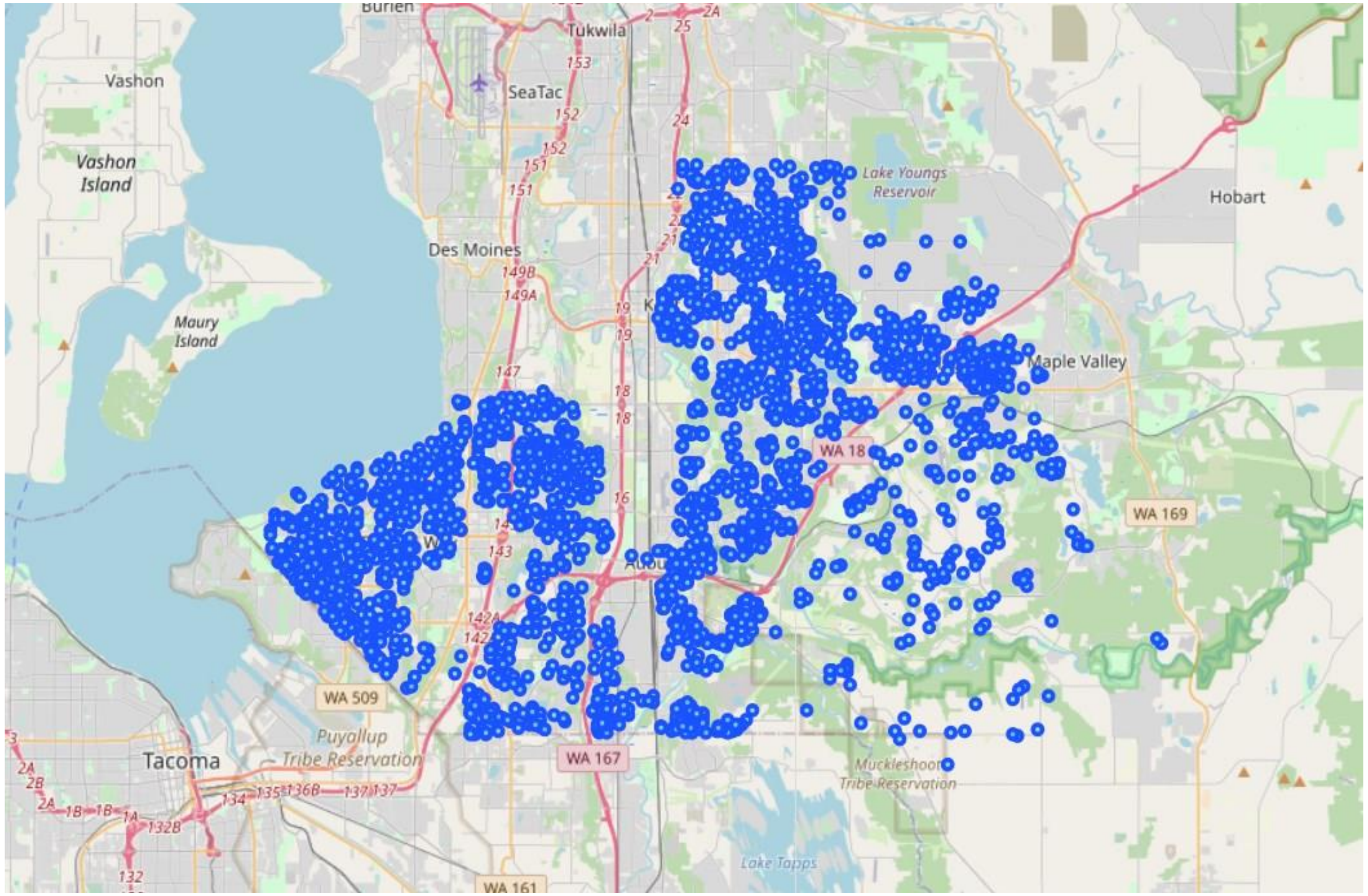
*More than 21.000 house records with features such as price, number of bedrooms, bathrooms etc.*

- Getzips.com

*Helped to classify the houses with area phone code, in order to select a portion of them to work with the models (i.e. houses with code 253)*

- Foursquare API

*Using the longitude and latitude of the houses, we extracted the top 30 popular venues in a radius of 500 from each house*




Distribution of houses (area code 253) in KC, Washington

# Data acquisition and manipulation

**At this point we have two data sets**

- One with the house records
- One with the venue records near the houses

**Data cleaning**

- Checked for any missing values (did not find any)
- Removed strange records (houses with no bedrooms or bathrooms)
- Created new feature (representing whether a house was renovated or not)
- Grouped the venues into broader categories (i.e. places to eat, attractions)
- Summarized each venue group per house (i.e. how many places to eat are near each house)
- Combined the two datasets

# Data acquisition and manipulation

**Feature selection**

- Discarded irrelevant features (id, date, zipcode etc.)
- Performed Pearson correlation to identify highly correlated features and discarded them

The final features that will be used to predict house prices are:

| | | | | |
|---|---|---|---|---|
| Bedrooms | Bathrooms | Total inter area | Total lot area | Floors |
| Waterfront | View | Condition | Total basem. area | Year build |
| Is renovated | Longitude | Latitude | *Places to eat* | *Places to drink* |
| *Places with sweets* | *Amusement* | *Facilities* | *Places for sports* | *Transport* |
| *Shops* | *Services* | *Beauty services* | *Attractions* | *Landmarks* |
| *Accomodation* | The features in *italics* represent the amount of venues of a specific type in the proximity of the house. | | | |

# Models for price prediction

**Machine Learning implementation**

- Split the data into two sets (75% - 25%, one for training and one for testing)
- Train the models based on the training set
- Use the features from the test set to predict a house price value
- Evaluate the predicted values by comparing them to the real house price values

# Models for price prediction

**Machine Learning algorithms used:**

- Linear
- Polynomial (degree of 2)
- Ridge linear
- Random forest
- Gradient boosting

# Models for price prediction

**Permutation importance**

We will identify the most important features per model

Then we will repeat the modeling with that features

**Model evaluation (Check the report for more details)**

- Root mean squared error (RMSE)
- Mean absolute error (MAE)
- R squared, both on training and test set
- Cross validation, splitting the data into four parts

# Models for price prediction

**The models created:**

*With permutation importance, we extracted the top features of each model, and tried it only with them. Therefore we have more than one model from each category*

- 3 linear (all features, top 15 and top 10)
- 4 polynomial (all features, top 350, top 300 and top 15 from linear model)
- 3 ridge linear regression (all features, top 15 and top 10)
- 6 random forest (with 50, 100, 200, 300, 400 and 500 trees each)
- 7 gradient boosting
  - 6 are combinations of number of trees (300, 400, 500) and learning rate (0.1 and 0.05),
  - the last was with top 21 features of 500 trees and learning rate of 0.05

# Results

**Linear and ridge linear models:**

- RMSE: ~ 50300 – 50800
- MAE: ~ 34100 – 34500
- $R^2$: ~ 0.75 and ~ 0.72 on cross validation
- Improved for the top 15 features

**Polynomial models**

- Due to many features, their scores are poor (RMSE > 54000, MAE > 35000)
- High $R^2$ on training set that suggest overfitting
- Improved when using only the top 15 features of linear model, but still not good enough

# Results

**Random forest**

- Better performance in general (RMSE ~ 48000, MAE ~ 31500)
- Greater values in $R^2$ test (~ 0.77) and cross validation (~ 0.74)
- Higher values in $R^2$ train (~ 0.95) that might suggest overfitting

**Gradient boosting**

- The best performing models (RMSE ~ 44000, MAE ~ 30500)
- Even better values in $R^2$ test (~ 0.8) compared to random forest
- More manageable values in $R^2$ train (~ 0.9)

# Results

## The best performing models per category

**Linear and ridge linear models:**

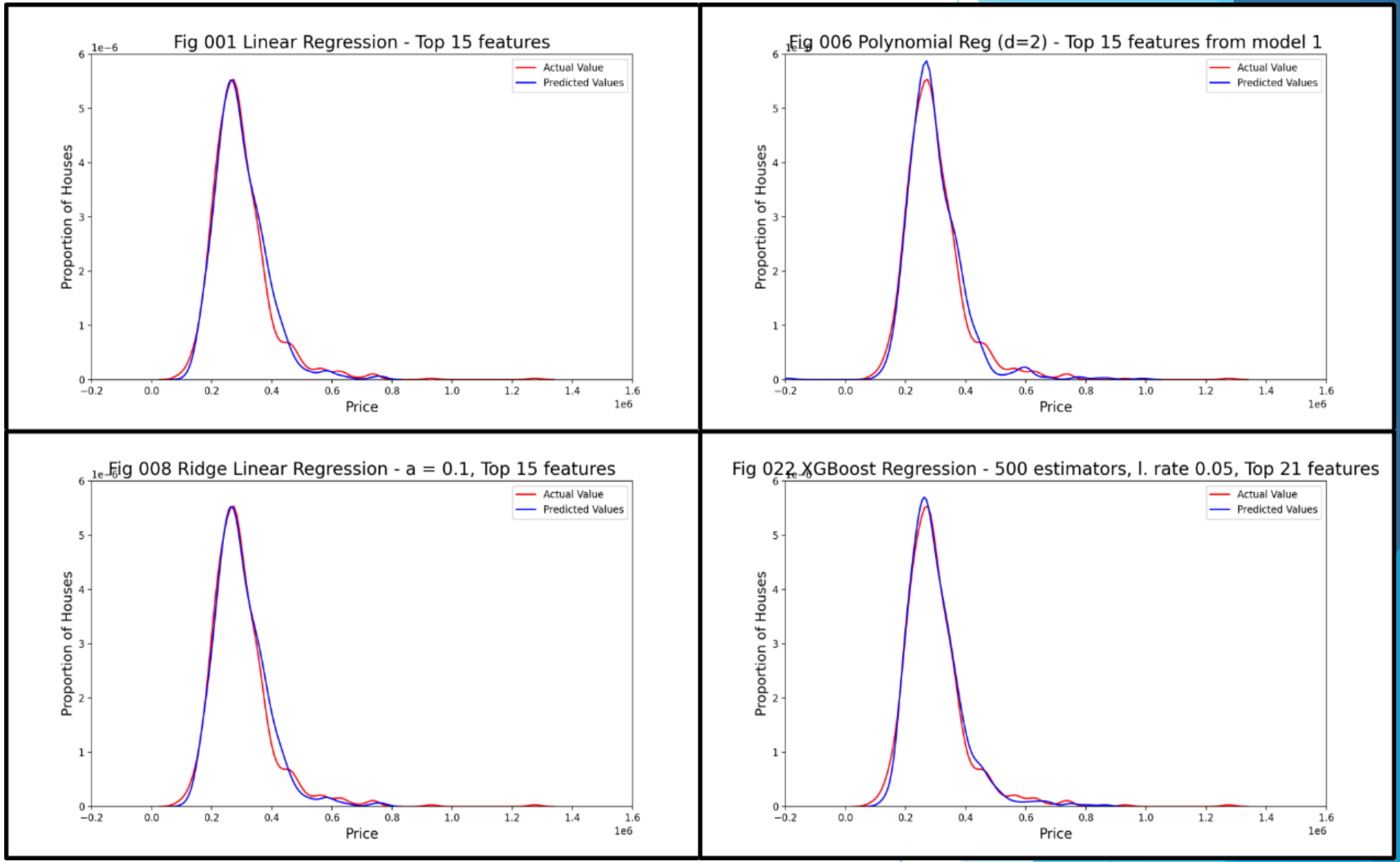
- Describe the peak of the distribution better
- Cannot predict houses with higher values that well

**Polynomial model:**

- Predicts way more houses with a price of 300000 (see the peak in the top right)
- Produces negative values

**Gradient boosting model:**

- Describes the peak well
- Describes values higher than 400000 better than the other models



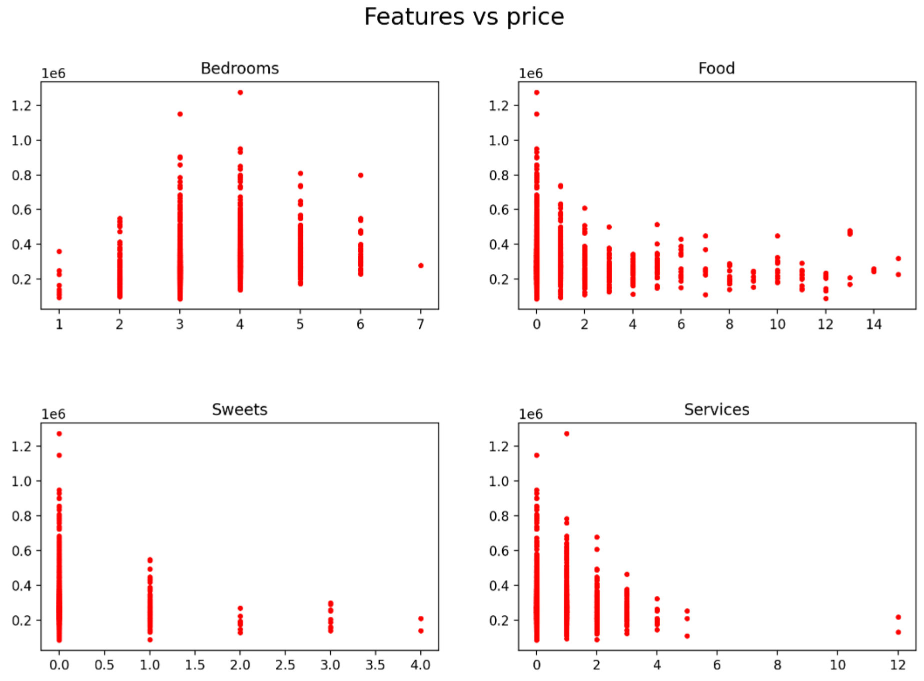
# Results

**Top features identified:**

- Total amount of living area was found among all models as impotant
- Lot and basement areas are also found important
- View was the most important feature in polynomial models
- The 13 features directly related to houses were found important in most models
- Food, sweet and service related venues were found important in linear models
- Shops and landmarks were found important in gradient boosting

# Discussion

- Tree-based models (random forest and gradient boosting) performed better
- House features had low linear relationship with house price (low Pearson correlation, as seen in the figure:



# Discussion

- Polynomial models did not perform well. But they can improve if fewer features are used
- The most important features are directly related to the house
- This analysis can be further improved if more data are available (e.g. type of heating, available garages or other spatial and temporal data such as elevation, mean temperature etc)

# Conclusion

- We created models in order to predict house prices
- We identified the most important features that drive the house price
- These findings can be utilized by real estate agents or individuals who want to sell/buy a house
- This results may apply in other areas with similar characteristics
- This methodology, given the right data, can be expanded for houses in an entire state or country
- Individuals who want to get the upper hand in real estate market must collect data and identify patterns among house features and their respective price

Thank you for your time!