King County House Price Predictions

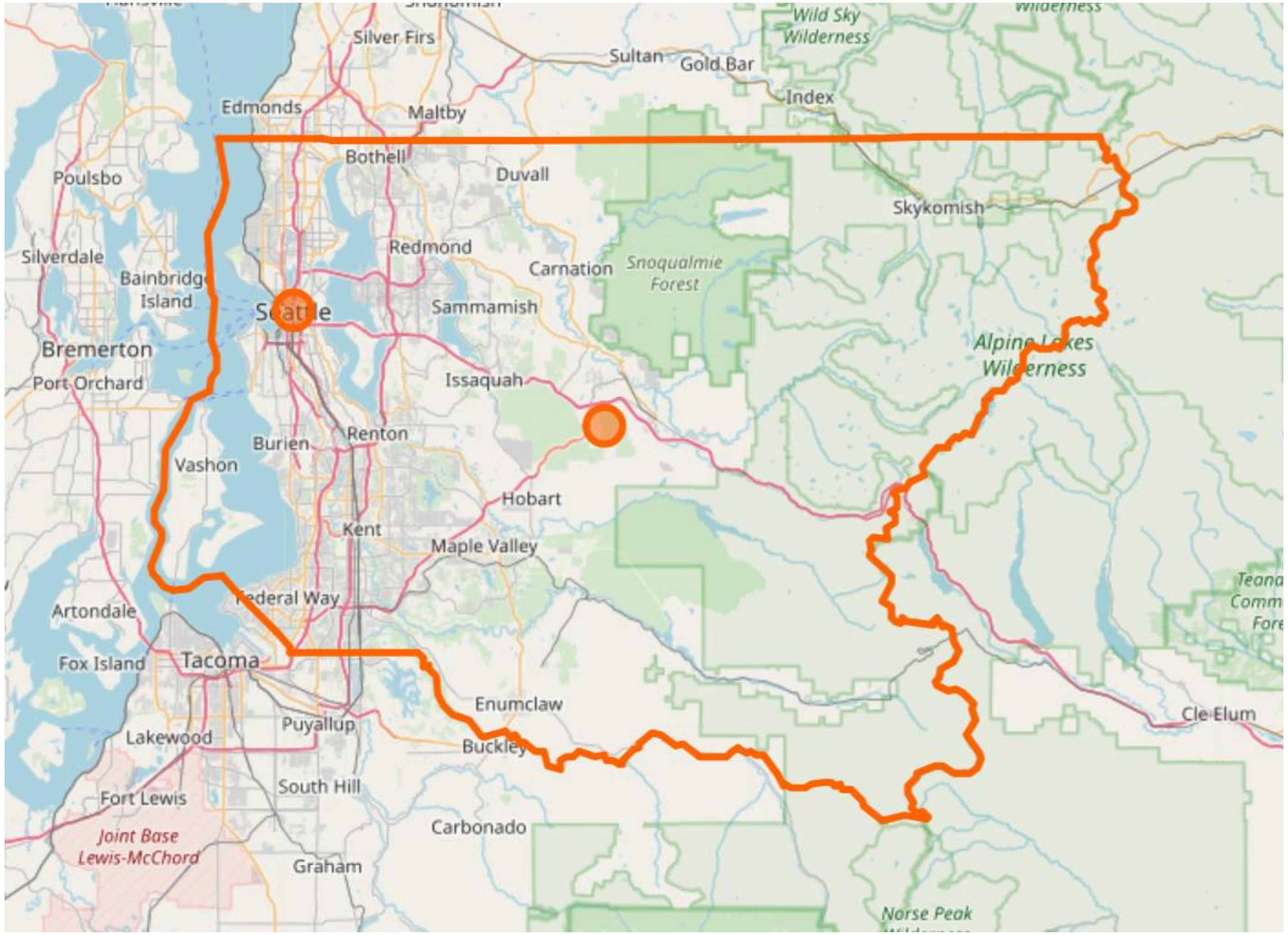
Springboard Data Science

Capstone project 1

Hye Joo Han

Project Goal

- Below (imaginary firm), a local real estate company serving King County, WA
- Goal: finding the best model(s) for house price predictions



Procedures

- Data collecting and wrangling
- Exploratory data analysis (EDA)
- Machine learning
- Final recommendation

Datasets

- Dataset 1: House Sales in King County, USA, Kaggle

  https://www.kaggle.com/harlfoxem/housesalesprediction
  - 21,613 house sales in King County
  - Sales between May 2014 and May 2015
  - 19 house features (bedroom, bathroom, square footage, year built, zip code, latitude and longitude, etc.), house id and sale price

- Dataset 2: Niche.com (https://www.niche.com)
  - Grades for public school, safety, cost of living, jobs, commute, etc. for each zip code
  - Collected by web scraping with a Python package, Beautiful Soup

Data Wrangling
House Sales in King County

- Removed 16 observations with zero bathroom or zero bedroom (0.07% of all rows)
- Made a new column *renovated* (1 for renovated houses and 0 for not) from the column for renovated years with 96% missing values
- Fixed a 33 bedrooms to 3 bedrooms which is more plausible
- Checked suspicious values using google map

Data Wrangling
Niche.com

- Collected grades (from A+ to D-) for 12 categories for each zip code (70 zip codes in total)
- 12 categories: public school, crime, cost of living, jobs, commute, nightlife, housing, good for families, diversity, weather, outdoor activities, and commute
- Removed crime which has the same grade for all zip codes
- Transformed the alphabet grades to score grades (from 4.3 to 0.7)
- Replaced missing diversity scores (for 4 zip codes) with median

Data Wrangling
Merge the two datasets

- Merged the second dataset (from niche.com) into the first dataset by left join on zip codes
- Now the new columns derived from zip codes are added to the main house sales dataset

Procedures

- Data collecting and wrangling
- **Exploratory data analysis (EDA)**
- Machine learning
- Final recommendation

Features correlated with house prices

House size related

sqft_living	.70
sqft_above	.61
sqft_living15	.59
bathrooms	.53
sqft_basement	.32
bedrooms	.32

Zip code related

good_for_families_score	.45
public_schools_score	.41
jobs_score	.33
cost_of_living_score	-.38

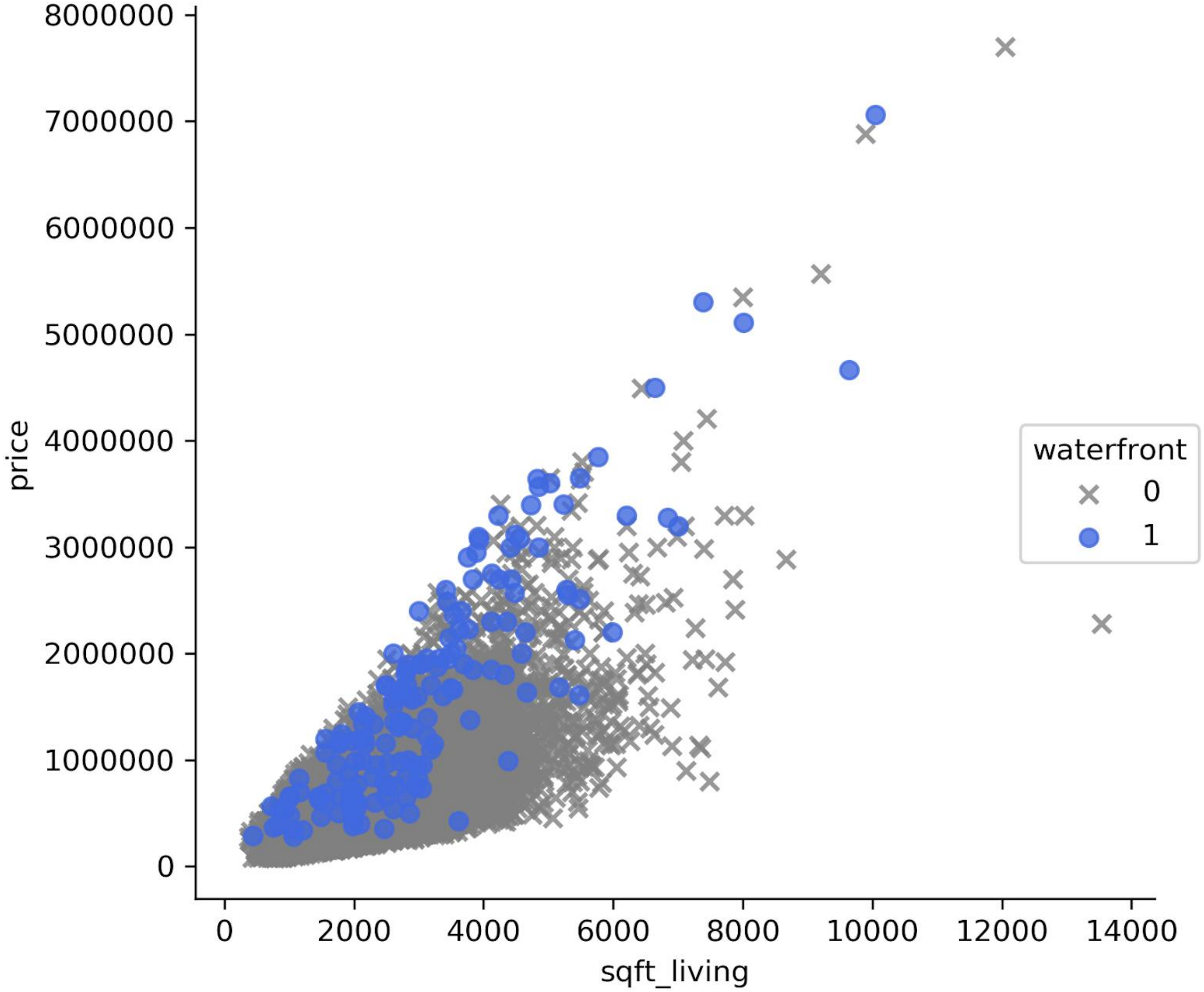
Others

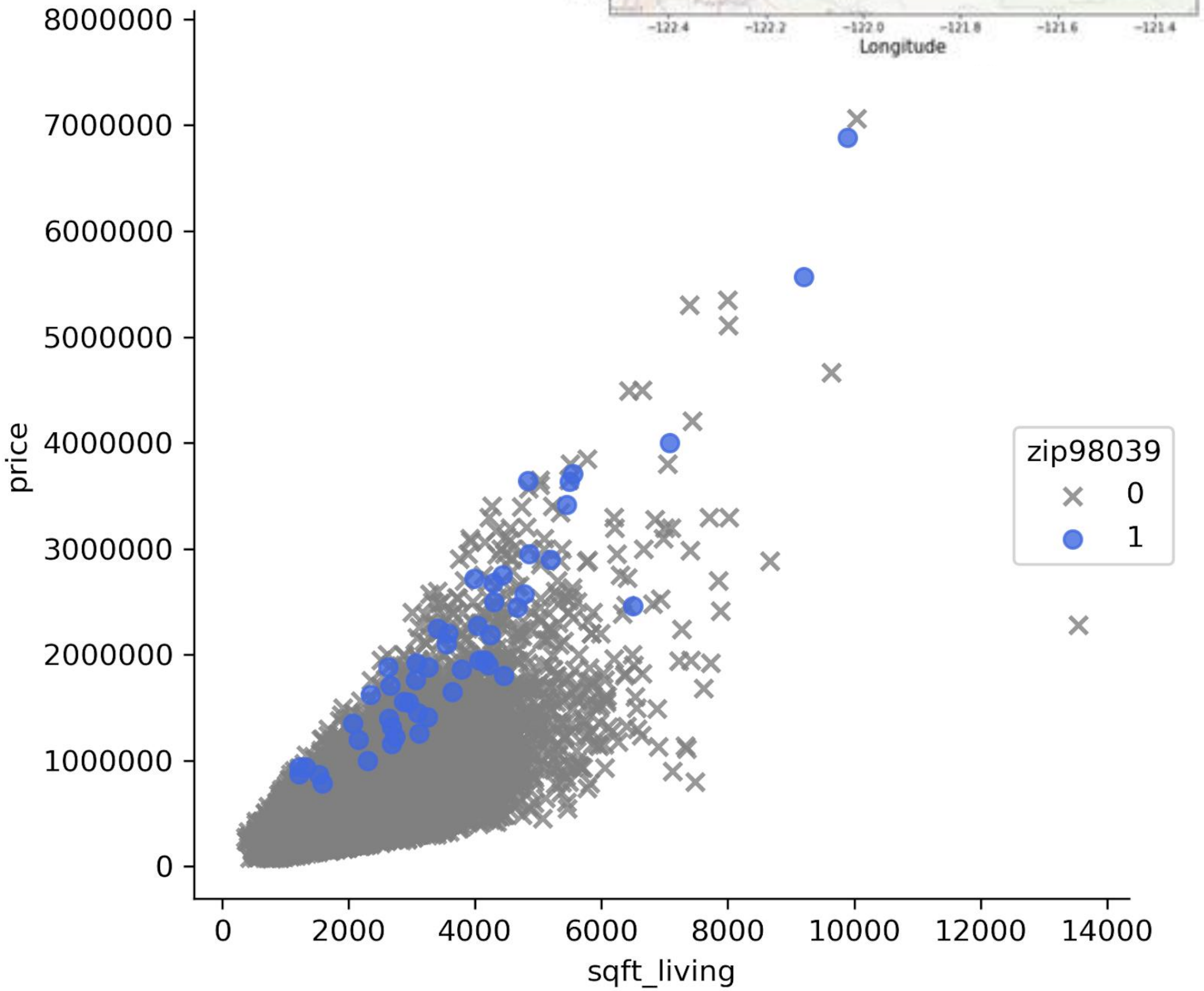
grade	.67
view	.40
Lat	.31

Strongly correlated independent variables

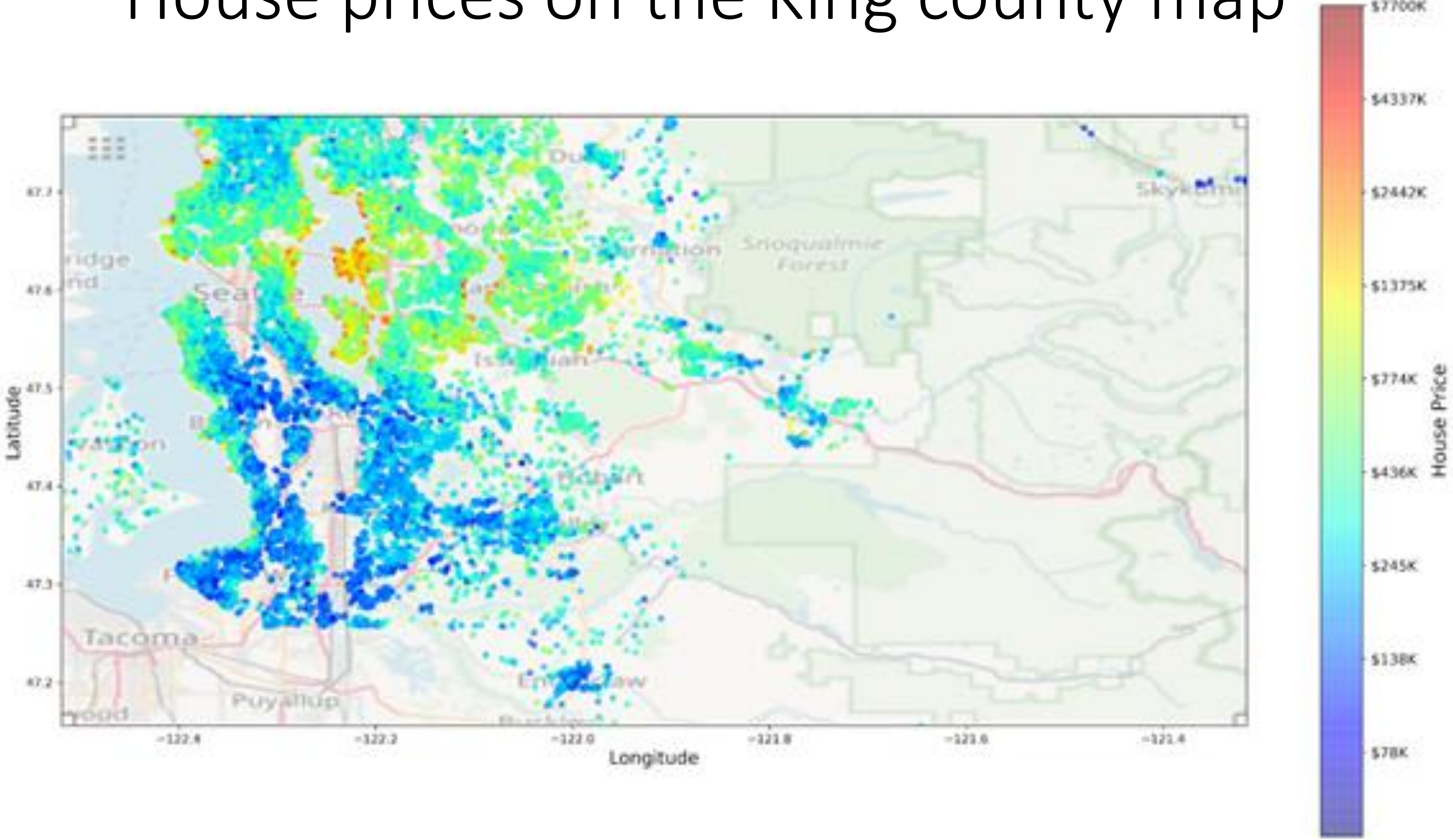
Good_for_families vs public_schools	.91
sqft_living vs sqft_above	.88
housing_score vs jobs_score	.77
sqft_living vs grade	.76
sqft_living vs sqft_living15	.76
grade vs sqft_above	.76
bathrooms vs sqft_living	.76
sqft_above vs sqft_living15	.73
sqft_lot vs sqft_lot15	.72
grade vs sqft_living15	.71
long vs weather_score	-.77

Waterfront and Zip code 98039

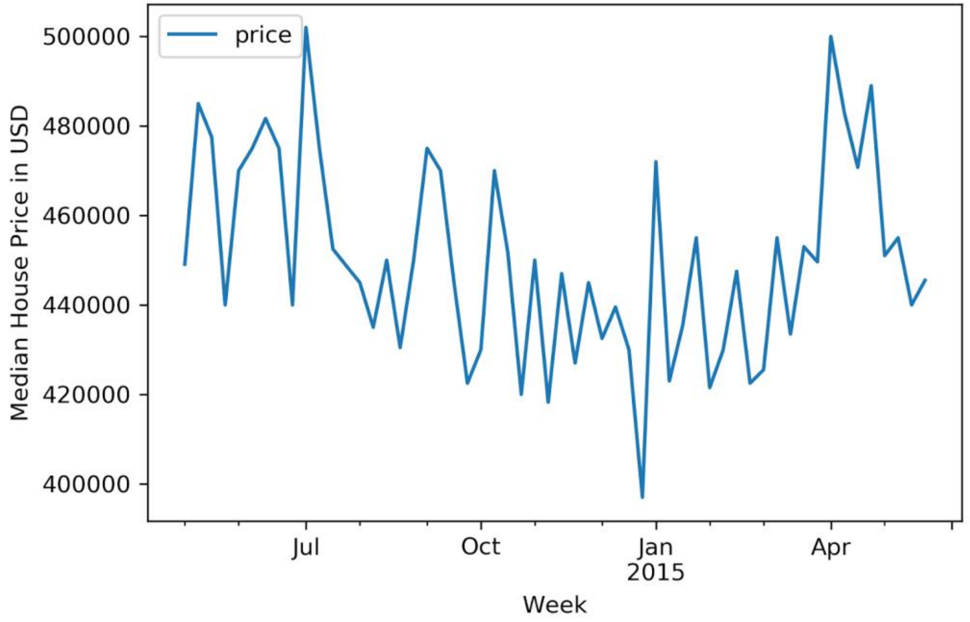


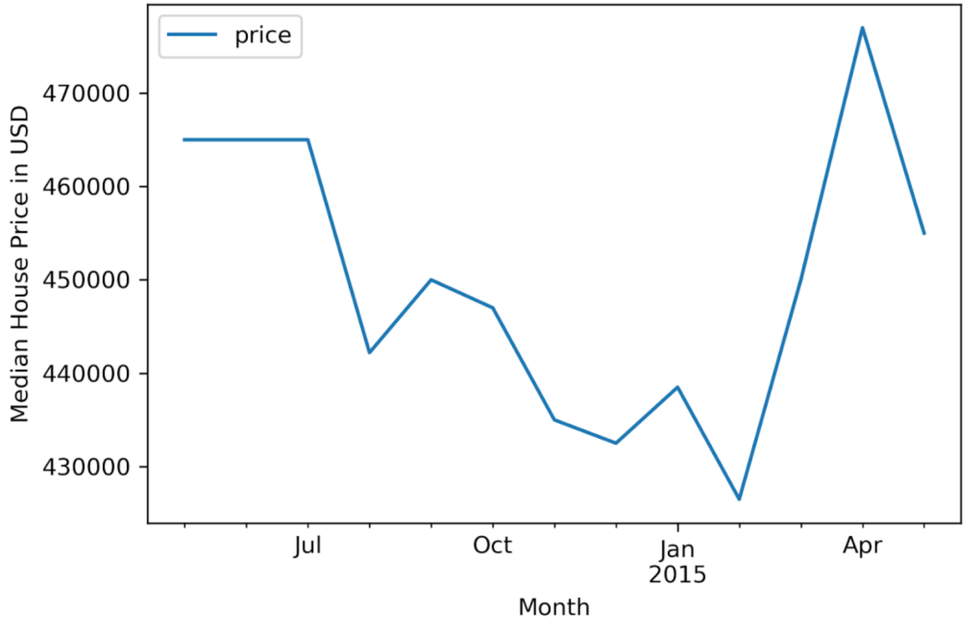


House prices on the King county map



Seasonal Fluctuations





- Highest between April and July and lowest around January or February
- New categorical variable sold_month to be made for machine learning

Procedures

- Data collecting and wrangling
- Exploratory data analysis (EDA)
- **Machine learning**
- Final recommendation

Data Preparation

- New features:
  - sold_month (extracted from sold dates)
  - renovated (1 for renovated houses 0 for not)
  - Zip98039 (1 for houses in 98039 0 for not)
- Removed features:
  - date
  - yr_renovated (96% missing possibly due to no renovation)
  - zipcode
- One-hot encoding for sold_month
- Test-training split

Algorithms

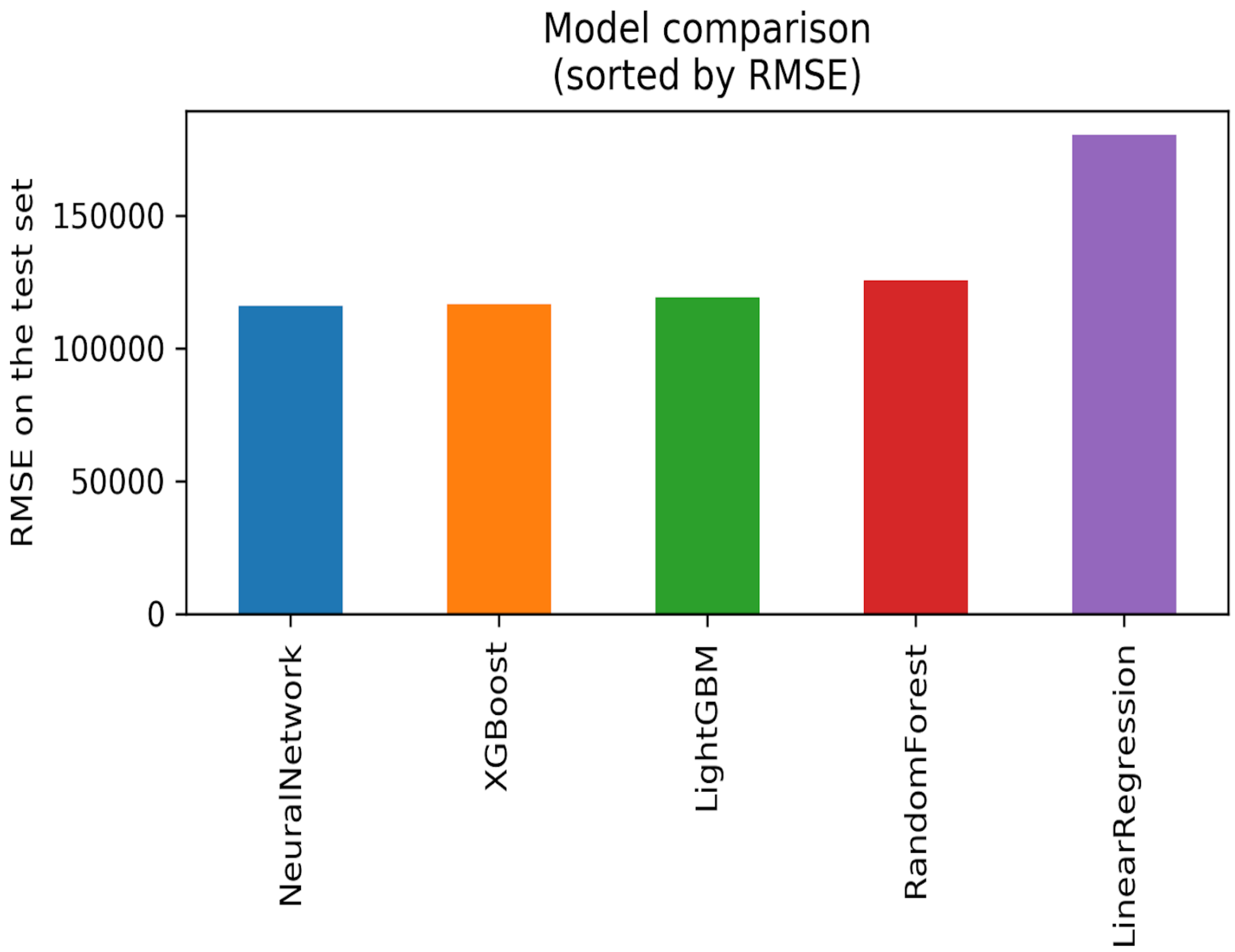
- Ridge regression (Linear regression)
- Random Forest (RF)
- XGBoost
- LightGBM
- Neural Network

Model Building

- Feature scaling
- 5-fold cross validation
- Hyperparameter tuning (coarse to finer)
- Early stopping for Neural network
- Metric: Root mean squared error (RMSE)
- R-squared as a reference

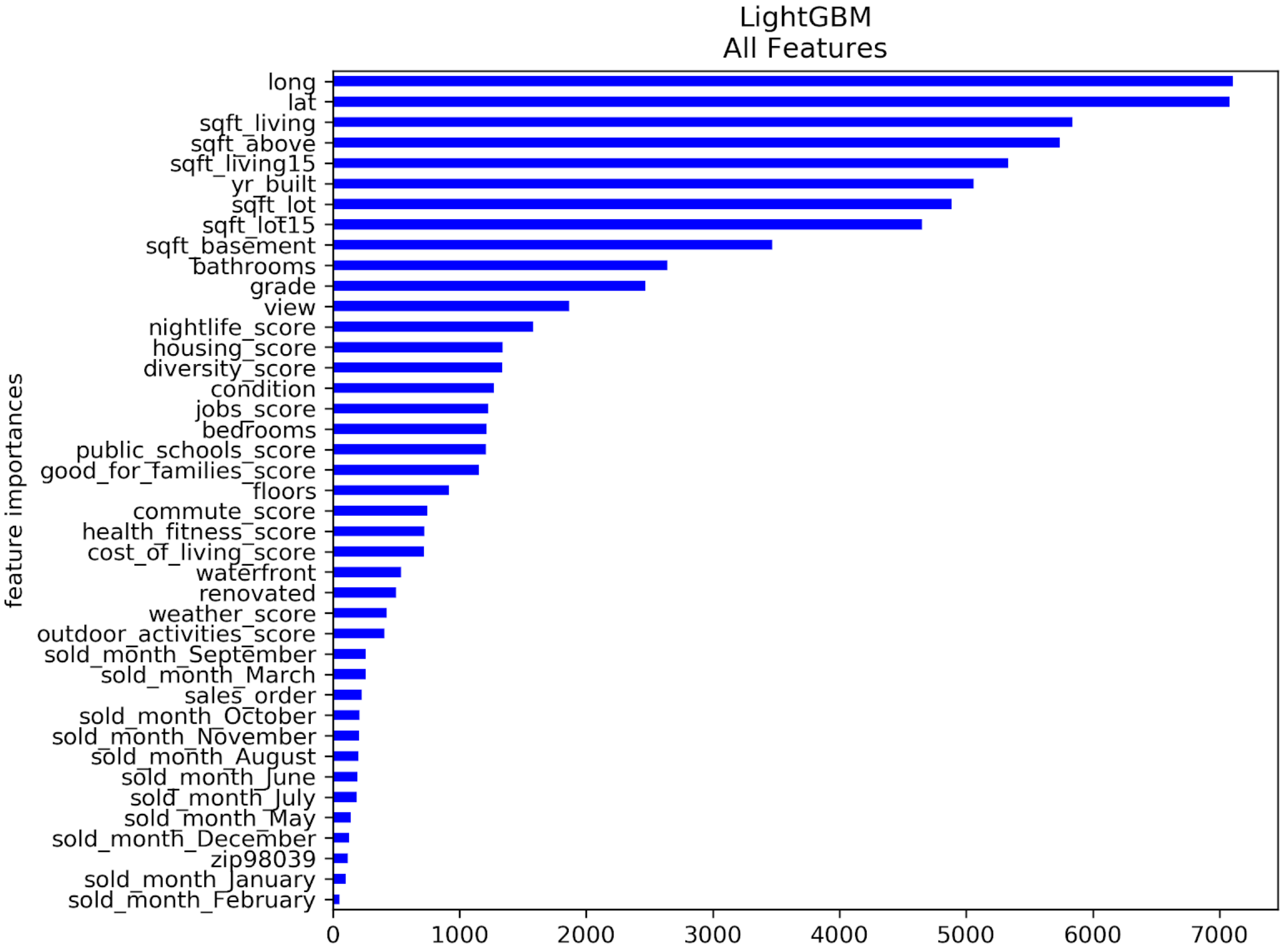
Performance

	RMSE	R^2	Time
NeuralNetwork	116252	0.896	23 min
XGBoost	116750	0.894	43 s
LightGBM	119439	0.890	19 s
RF	125872	0.878	3 min
Linear Reg	180464	0.750	507 ms



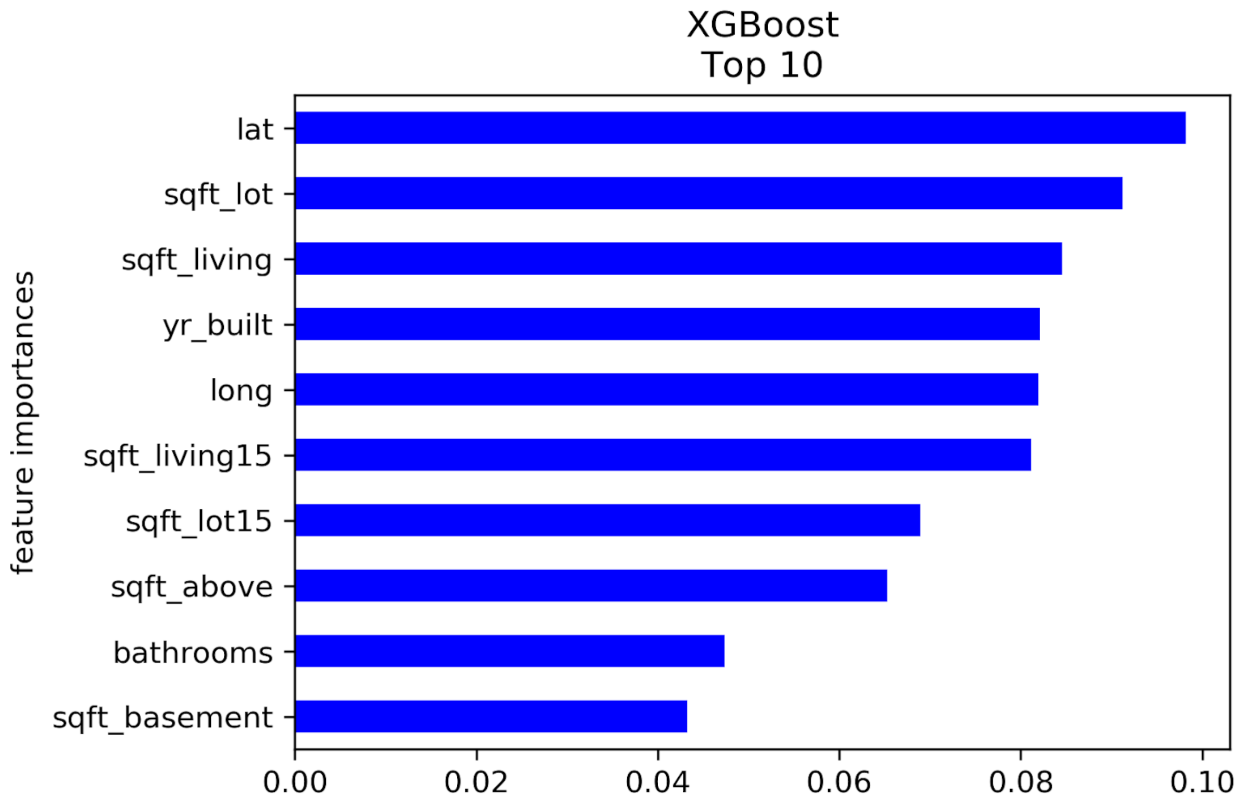
- Winners: Neural network, XGBoost and LightGBM (similar RMSE)
- XGBosot and LightGBM are more reliable, fast, and convenient (feature_importances_)

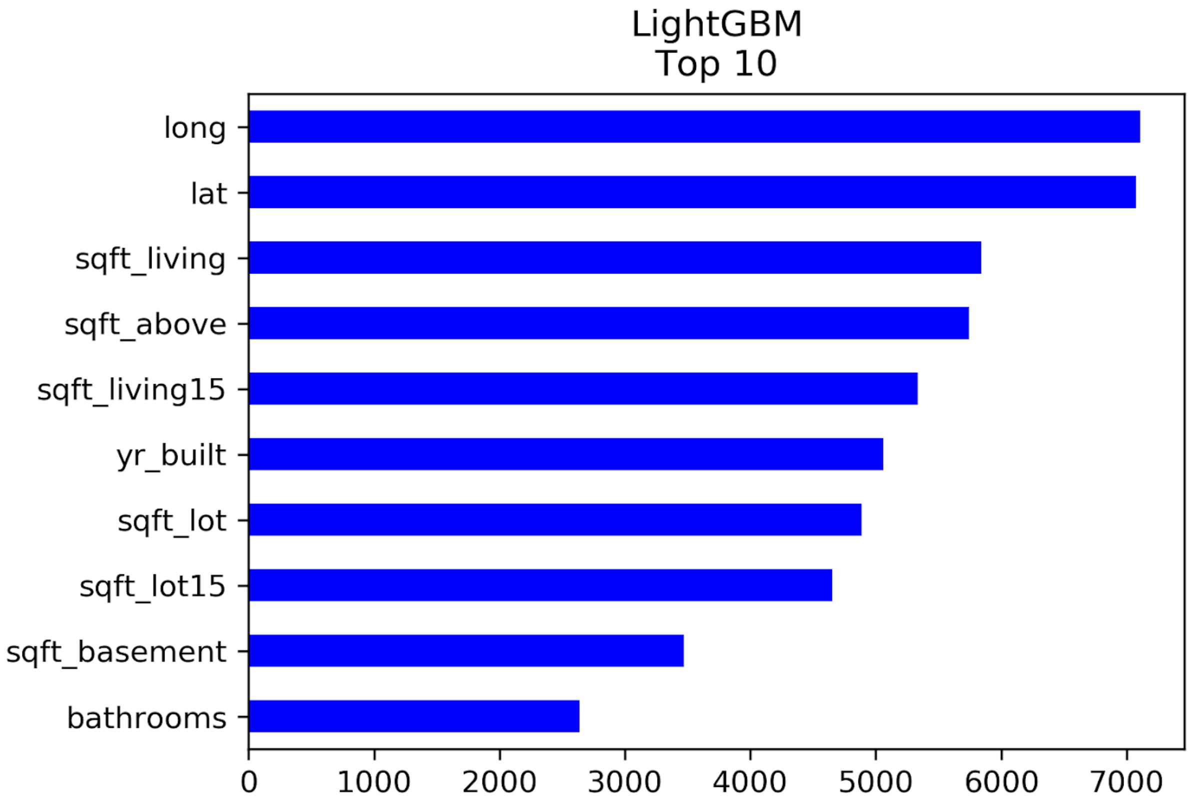
Feature Importances (All)



- XGBoost and LightGBM approximately agree the order of feature importances (cf. RF)
- Moderately important features: those related to environment scores
- Least important features: those related to sold months
- All features contributed to some extent (cf. RF)

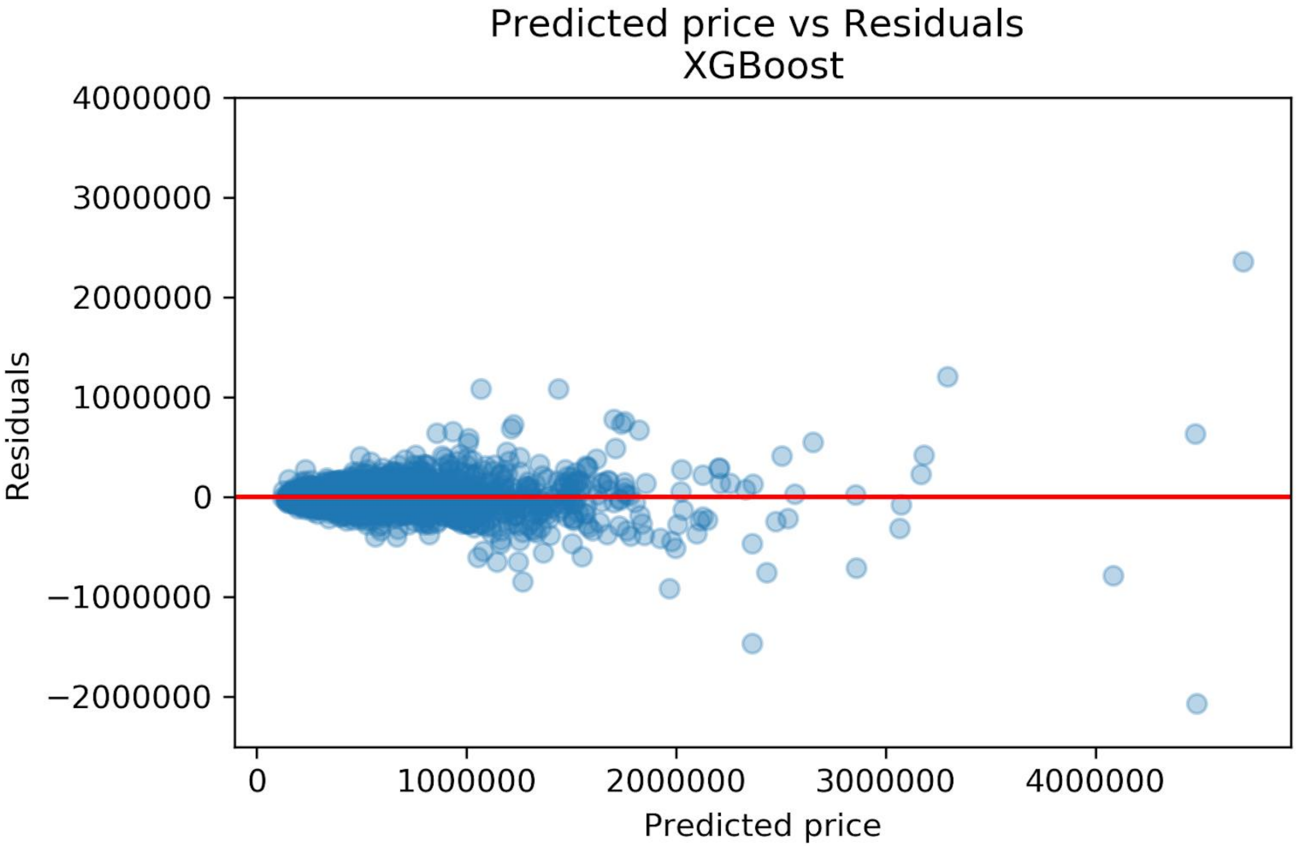
Feature Importances (Top 10)

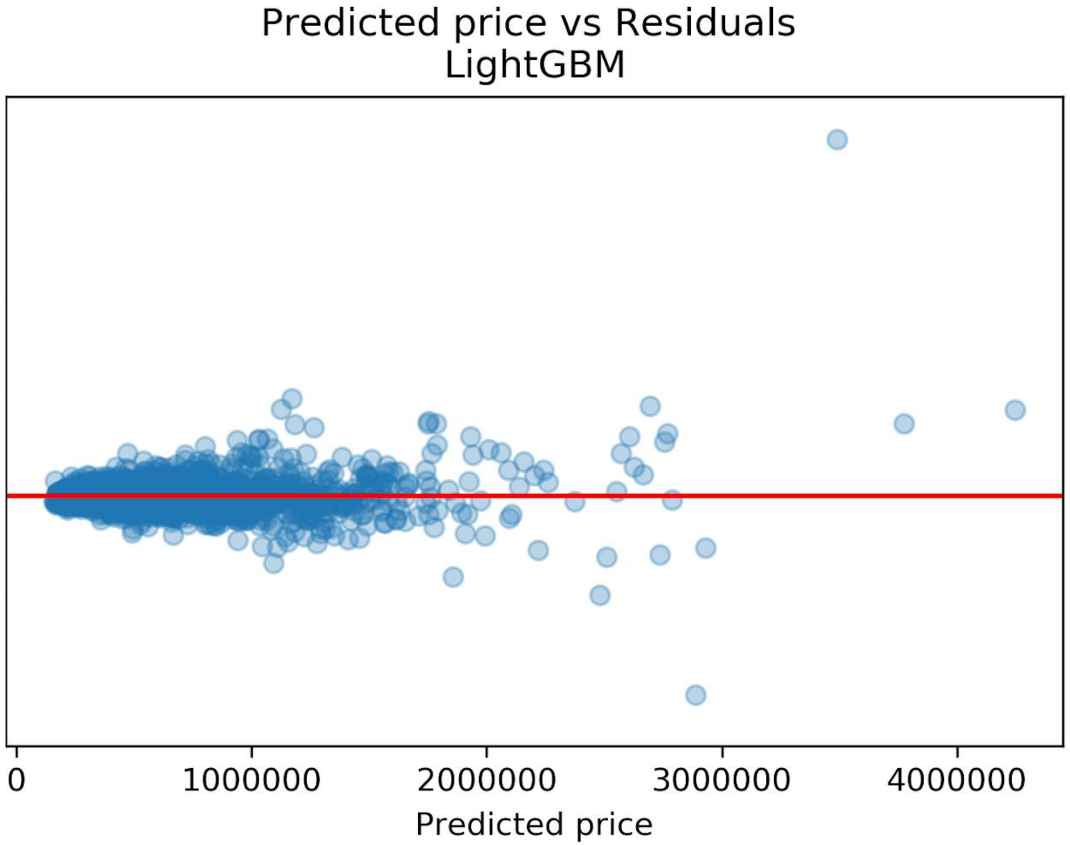




- Exactly same top 10 features:
  - latitude and longitude
  - square footage related features
  - built years and number of bathrooms

Some Error Analysis



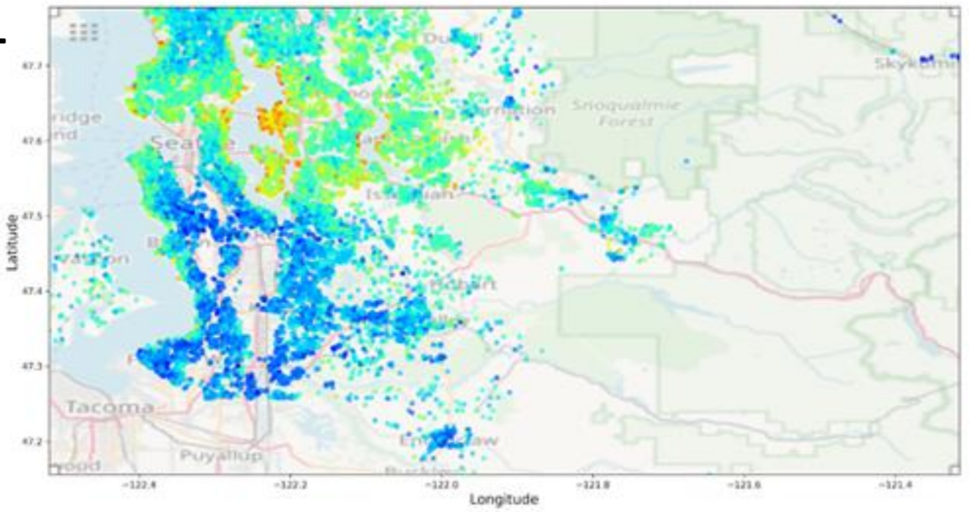


- XGBoost makes more outliers with residuals over 1 million or less than -1 million than LightGBM (6 vs. 2)
- LightGBM model has one extremely big residual over 3 millions (waterfront)

Final Recommendation

- Important features suggested by both EDA and ML
  - square footages of living area
  - number of bathrooms
  - latitudes

  (positively correlated with house prices)



- Important, but not linearly related to house prices
  - longitude
  - year built
- Better to buy a house in winter and sell a house around late spring or early summer price-wise
- Suggest the XGBoost or LightGBM models (high speed and low RMSE)
- Carefully determine price for a house with extremely high predicted price (say, over 2 millions)

Links

- Final report

https://docs.google.com/document/d/15UNqqwrmXJjWTMq_q4ewASS5IWYUKGvQ8-Gdojg4t_0/edit?usp=sharing

- Jupyter notebooks on Github

https://github.com/math470/Springboard_Capstone_Project_1