PROJECT LUTHER SUMMARY
PREDICTING PRICE OF HOUSE IN KING COUNTY

-Hiranya Krishna Kumar

DOMAIN

In Project Luther, I predicted house prices in King county by scraping data from Realtor.com and bestplaces.net using selinium and beautiful soup to extract features such as zipcode, school-rating, year built, home-type, sqft-living, sqft lot, year built , View type etc from realtor.com. Furthermore, I scraped data for median age, median income, commute-time by Zipcode from bestplaces.net. My final data-set after dropping nan's contained 1382 data points and 15 features used for modeling .

DATA

Variable	**Type**	**Description**	**Used for model**
Price	int	Price of house	Target
Median Age	float	Median age in zip code	Y
Bedrooms	float	No of beds	Y
Bathrooms	float	No of baths	Y
Sqft living	int		Y
Sqft Lot	int		Y
High School Rating	float	Rating of schools	Y
Middle School Rating	float		Y
Elementary School Rating	float		Y
Median Income	float	Median income in zip	Y
Zipcode	str		N
Year Built /home Age	str/int		Y
Commute time	int	One way commute time in a zip code	Y

PROJECT LUTHER SUMMARY
PREDICTING PRICE OF HOUSE IN KING COUNTY

-Hiranya Krishna Kumar

Home Type	str	Single family/ Condo/ townhome/ apartment	Y
City	str		N
Fireplace	float		Y
Heating	str/binary	Heating type	N
Cooling		A/C or ceiling fan	Y
View type	str/binary	waterfront/non waterfront	Y

DATA CLEANING AND FEATURE ENGINEERING

After data collection, I replaced all missing values with 0 for categorical variables and used median imputation for numeric variables like Home Age,bedrooms, bathrooms. All categorical variables were consolidated into two categories and hot encoded. For all condos sqft_lot was imputed as 0. Exploratory data analysis was done on certain categorical variables like cooling, View type and it was found that homes with cooling and view had higher median price .Home prices varied significantly by zip-code with Medina and Bellevue having the highest median price by zip-code.
The histogram of target variable , price was skewed to the right so I applied a log transform to normalize the distribution of the target variable. I also looked at the heatmap of the correlation of different features and them to have low correlation with each other except fro school ratings. No of bedrooms, school ratings, bathrooms, Median income had the highest correlation with Logprice.

MODEL BUILDING

An MVP using stats model with all the 15 features had an adjusted R2 of 0.565 . Removing collinear features (coefficients that were not significant

PROJECT LUTHER SUMMARY
PREDICTING PRICE OF HOUSE IN KING COUNTY

-Hiranya Krishna Kumar

p-value>0.05) improved the adjusted R2 to 0.568 , so I kept all the features for further model tuning.The data was split into train and test set using a 70:30 ratio. Since the dataset had varying orders of magnitude standard saler was applied to normalize the dataset. On the Train set polynomial feature transformation was done on the features varying the degree from 2-6 using a 10-fold cross validation and found degree 2 to have highest R2 of 0.664.

Lasso and Ridge regularization was applied using different orders of magnitude of alpha for both the regularizations with 3-fold cross validation.Lasso regularization with 2nd degree polynomial features and 3-fold CV had a higher R2 with 0.78 at an alpha of 10^{-5}

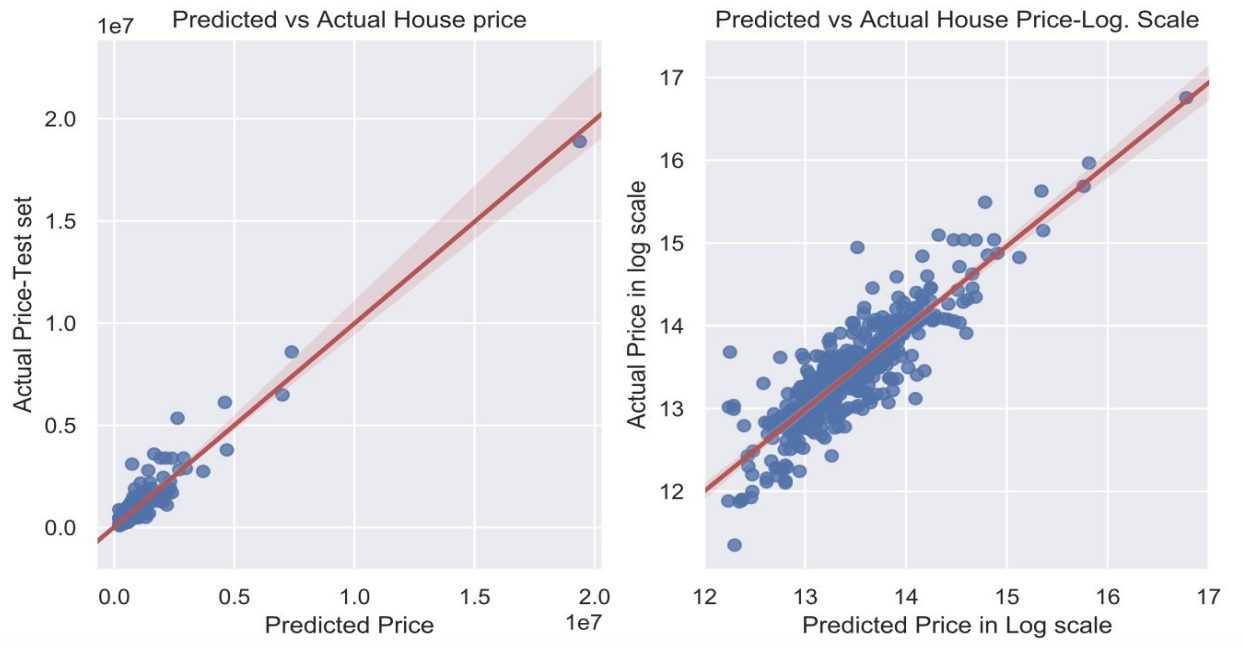
RESULTS

The final model using Lasso regularization was applied to the test set. The results are tabulated below

PARAMETERS	**VALUE**
TEST RMSE(e^(sqrt(MSE)) % error between predicted and actual prices=5.5%	1.055
TEST R-Square	0.751
DEVIATION IN MSE BETWEEN TEST AND TRAIN SETS	2.67%

PROJECT LUTHER SUMMARY
PREDICTING PRICE OF HOUSE IN KING COUNTY

-Hiranya Krishna Kumar



FUTURE WORK

1. Add more features like crime data by Zip-code, Condition of house
2. Use Zip-code as a ordinal categorical feature
3. Categorize features by Type i.e. Condo/ Townhouse/ Single-Family/ Apartment rather than only single-family or non single family
4. Get more data for houses with prices between 2 million-20 million
5. Use median imputation by zip-code to impute missing values