1. Introduction

The real estate sector is an important industry with many stakeholders ranging from regulatory bodies to private companies and investors. Among these stakeholders, there is a high demand for a better understanding of the industry operational mechanism and driving factors. Today there is a large amount of data available on relevant statistics as well as on additional contextual factors, and it is natural to try to make use of these in order to improve our understanding of the industry. Notably, this has been done in Zillow's Zestimate [4] and Kaggle's competitions on housing prices [2]. In some cases, non-traditional variables have proved to be useful predictors of real estate trends. For example, in [3] it is observed that Seattle apartments close to specialty food stores such as Whole Foods experienced a higher increase in value than average. This project can be considered as a further step towards more evidence-based decision making for the benefit of these stakeholders. The project focused on assessment value for residential properties in Calgary between 2017-2020 based on data from [1]. The aim of our project was to build a predictive model for change in house prices in the year 2021 based on certain time and geography dependent variables. The main steps in our research were the following. •

Exploratory Data Analysis (EDA).

By conducting explanatory data analysis, we obtain a better understanding of our data. This yields insights that can be helpful later when building a model, as well as insights that are independently interesting.

Feature Selection

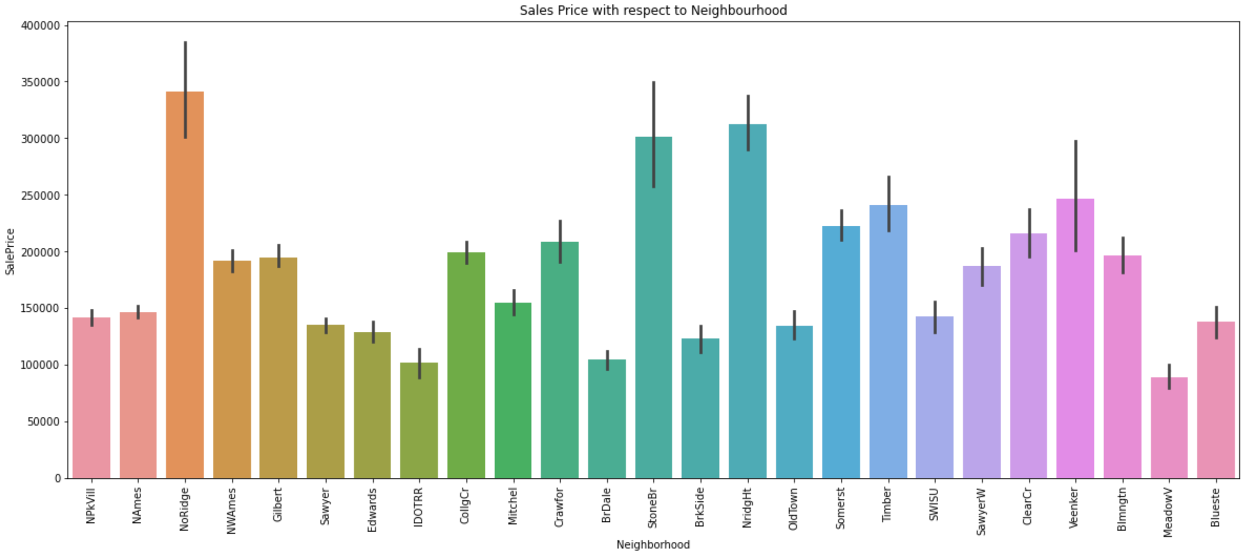
In order to avoid overfitting issues, we select 20(according to PCA) variables out of the original 36 by using methods LASSO , elastic net , forward feature selection, backward feature selection.

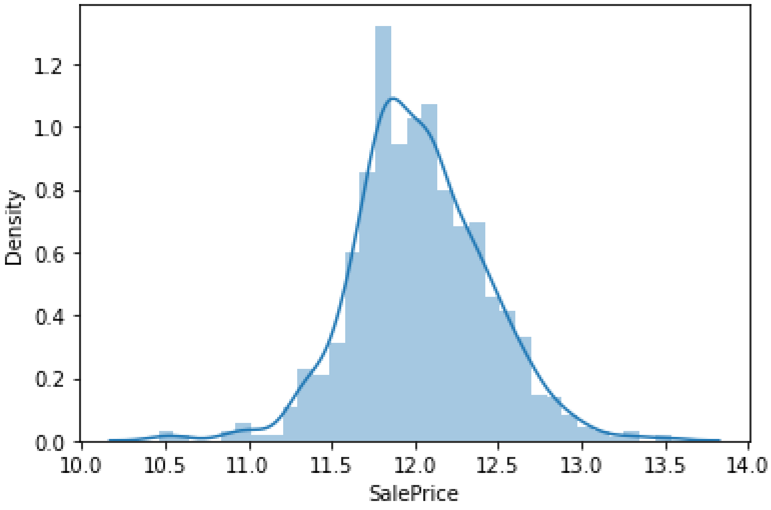
Modeling

We apply Decision Tree , Random Forest models for prediction of the percentage change of the housing prices.

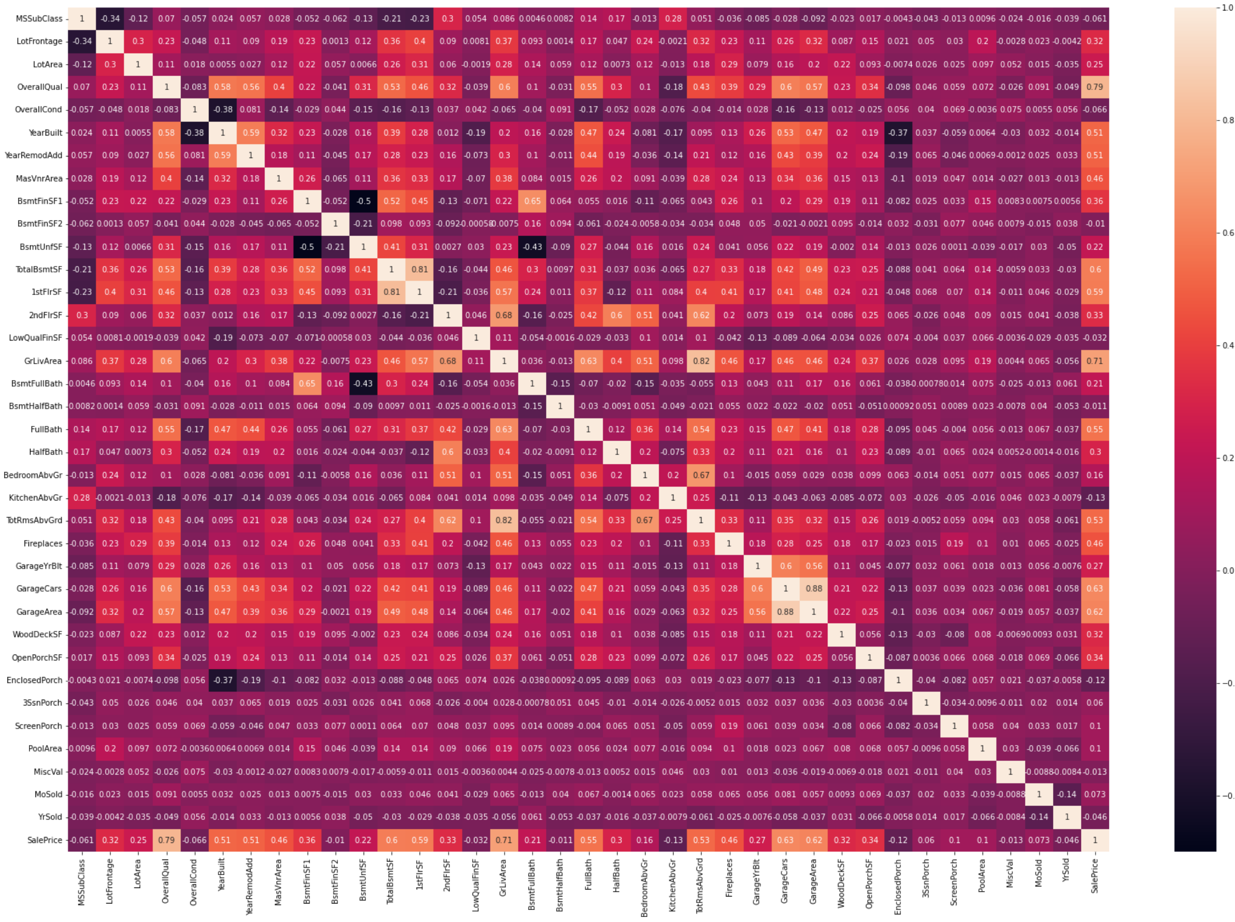
Exploration of reasons for misclassification in model We then go back to the original data to find out why some samples are misclassified by our model. In this report, we describe our approach to these steps and the results that we obtained.

2. Exploratory Data Analysis Figures.





As part of the EDA, we first looked at the mean percent change of the housing prices from 2017-2020 for each FSA whose data is given. Figure 2 suggests that on HOUSING PRICE PROJECT REPORT an average there was positive change in prices in the Year 2018. In order to analyse our features more carefully, we also looked at the correlation of various features of the houses.



Correlation of Features

Figure gives us an insight on how parameters are correlated with each other.

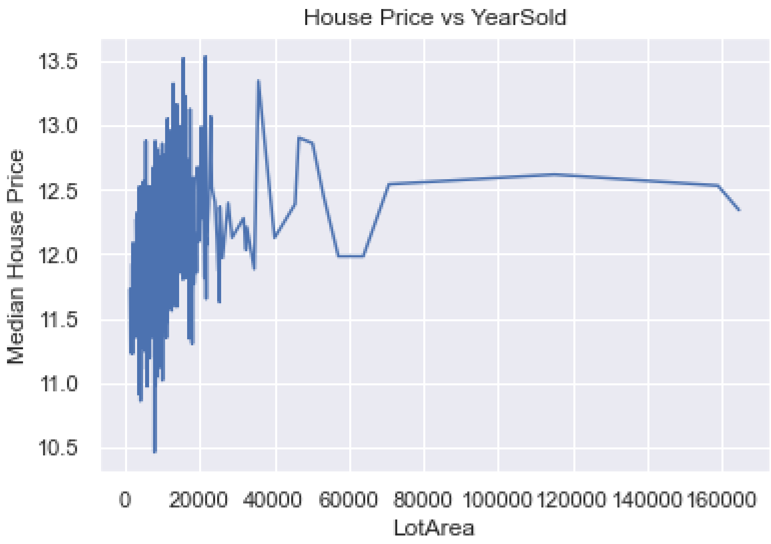
In order to understand our data, we first perform exploratory data analysis. This will provide us with insights that will be useful in building prediction models, as well as insights that may be of interest to stakeholders. As part of the Exploratory Data Analysis we aim to: • Look into the relationship between each variables and annual house price percentage change, and identify any patterns. For example, between the year of construction of a house and its annual percent price change. • We will also analyse relationships between the features. This may reveal that certain features are redundant and this would help the subsequent analysis.

Methodology

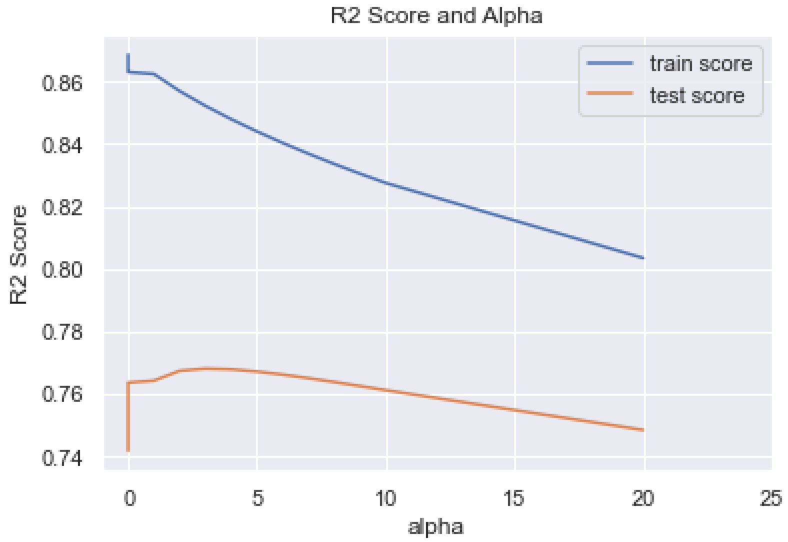
Feature selection. Our data has 36 features in total. If we use all of them in our prediction model, the model will have a risk of overfitting. Therefore, we decide to remove some unimportant features. We choose a dimensionality reduction algorithm called Principal Component Analysis (PCA) as the method to estimate how many components are needed to describe the data. The optimal number of features for the prediction can be determined by looking at the cumulative explained variance ratio as a function of the number of components. This curve quantifies how much of the total, 36-dimensional variance is contained within the first n components. For example, we see that with the digits the first 10 components contain approximately 90% of the variance, while you need around 25 components to describe close to 100% of the variance. Here we see that our

House-Price-Prediction Assignment

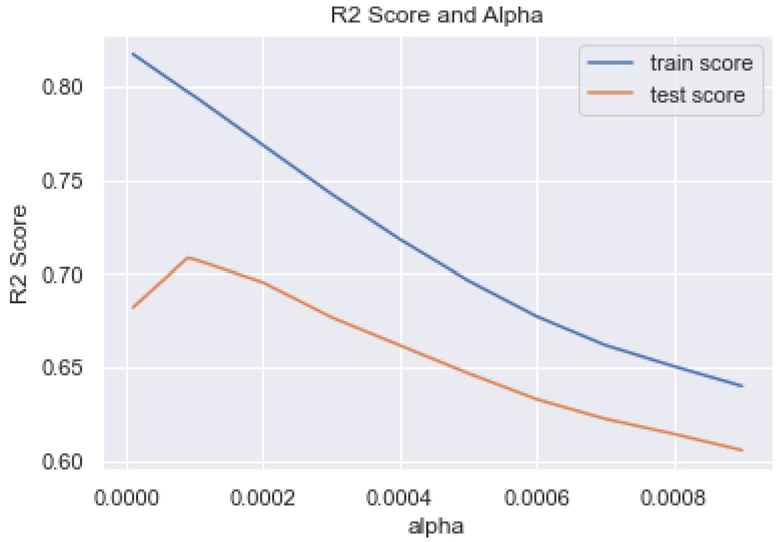
Question 1: What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented? Answer: In the case of ridge regression, when we plot the curve between negative mean absolute error and alpha, we see that as the value of alpha increase from 0 the error term decreases, and the train error is showing increasing trend when value of alpha increases. when the value of alpha is 2 the test error is minimum, so we decided to go with value of alpha equal to 2 for our ridge regression. For lasso regression I have decided to keep very small value that is 0.01, when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value zero. Initially it came as 0.4 in negative mean absolute error and alpha. When we double the value of alpha for our ridge regression no we will take the value of alpha equal to 10 the model will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and no thinking to fit every data of the data set. From the graph we can see that when alpha is 10 we get more error for both test and train. Similarly, when we increase the value of alpha for lasso, we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases. The most important variable after the changes has been implemented for ridge regression are as follows: - 1. MSZoning_FV 2. MSZoning_RL 3. Neighborhood_Crawfor 4. MSZoning_RH 5. MSZoning_RM 6. SaleCondition_Partial 7. Neighborhood_StoneBr 8. GrLivArea 9. SaleCondition_Normal 10. Exterior1st_BrkFace The most important variable after the changes has been implemented for lasso regression are as follows:- 1. GrLivArea 2. OverallQual 3. OverallCond 4. TotalBsmtSF 5. BsmtFinSF1 6. GarageArea 7. Fireplaces 8. LotArea 9. LotArea 10. LotFrontage

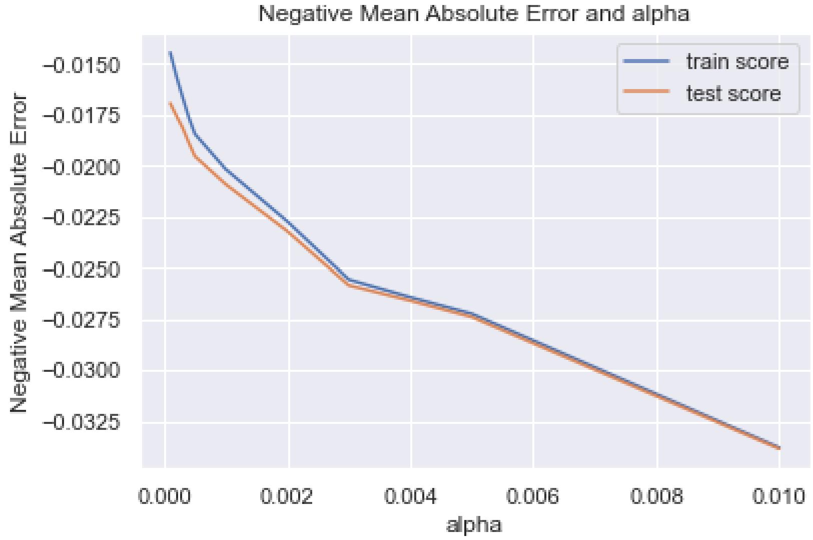


Question 2: You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why? Answer: It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance and making the model interpretably. Ridge regression uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Residual sum or squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression. Lasso regression uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.



Question 3: After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now? Answer: Below are five most important predictor variables that will be excluded are :- 1. GrLivArea 2. OverallQual 3. OverallCond 4. TotalBsmtSF 5. GarageArea Question 4: How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why? Answer: The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the BiasVariance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data. Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data. Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.





Name- Promila
Organization-Flip Robo