Project-2 Check-in

Any surprises from your domain from these data?

I spent a good amount of time looking at the data and understanding the business domain last week itself before selecting the project. I had some understanding of the business as I bought a house in last few years. I started this project with few assumptions in mind of what are the factors which are going to be most important. However, data is showing the correlation somewhat different than what I thought. Overall condition (which I thought is vague data) is playing more important role than basement condition. Given the current scenario of real state market, there is a lot of other factors impacting the final price, including lumber price, logistics issues and labor market and available labor force in the housing area. A new house construction is costing 1.25 to 1.5 times more given the above stated market conditions. This project will train a model with the data available at hand but will need to review how it does in current market and how can be accommodating of those factors as market is constantly changing.

The dataset is what you thought it was?

Data set has good information around the attributes of a house and the final sale price. However, it has 43 categorical variables which I need to encode for my regression algorithms. A lot of these categorical variables have weak correlation with the target variable sale price. Attributes like basement condition, basement type, garage quality are not correlated at all (under 0.10) with sales price. I still need to encode them and then apply some feature selection technique to filter out the attributes which are not going to help train the model. There are a few attributes which are strongly correlated inversely, which means, when the value of those attributes increases the sale price reduces. There are a lot of numerical attributes too, however, many of those have strong correlation with the sale price. There is a few numerical attributed needing filtered out given the weak correlation.

Have you had to adjust your approach or research questions?

Nothing in the data and business domain has made me think that I need to adjust my approach yet. I trained a basic random forest regressor model by the time of writing this milestone and it is giving R2 score of around 89%. I have trained a few other models too like Gradient boosting regressor, Linear regressor and a few more and getting decent scores from all of those.

The research questions I have in this project is the ability to fairly predict a sales price for a house give house attributes. It looks like those still hold valid and I will be able to come up with a good model to use to answer those questions.

Is your method working?

Yes, I have done a few regression models so far in this program. I almost follow the similar steps every time and it works with a few deviations. One thing I might need to work on is, using the correct encoder. I am using label encoder and I sure that is not the best way to encode all the categorical variables. I have seen fancy encoding techniques that train a model first to choose the correct encoding technique.

I tend to use the algorithms I have used in the past and do not look out for newer or other algorithms available. I sure want to try if time permits. However, I make sure to do exhaustive training and hyperparameter tuning. This helps me get better results. Apart from these two concerns, I am not planning to change anything from my plan so far.

What challenges are you having?

The machine I use has limited resources and access. Anytime I want to download any new package, it stops me sometimes, and it works other times. I bought a better machine and will spend some time on it to set it up.

I want to write code once and use it again for later. When I do data cleanup on my train data, I want those steps to be a part of a pipeline which I can run through with the test data and eventually to the live data when model is deployed in the production. I have not been successful doing that in this course yet, but might try that later.