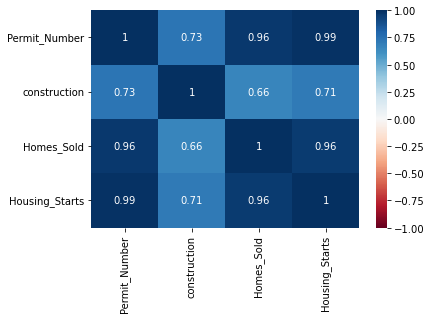
**Building a Data Science model to find the factors which influenced U.S home prices the most in the last 20 years.**

**ABOUT THE DATASET**

* **The dataset contains key factors that could influence Residential home prices in the last 20 years in the United States. This factor falls into two categories**
* **i.e. Supply & Demand**
* **The S&P Case-Shiller Housing Price Index (HPI) is taken as the y variable, or dependent variable, as an indicator of change in prices.**
* **Supply dataset (Monthly data)**
* **Building Permits (Permit Number)-Number of building permits allotted**
* **Construction Spending (Million $)-The amount spent (in millions of USD) is a measure of the activity in the construction industry.**
* **Housing Starts (New Housing Project)-This is a measure of the number of units of new housing projects started in a given period.**
* **Homes Sold (units)-House for sale is a basic measure of supply.**
* **Demand dataset (Quarterly data)**
* **Mortgage Rates (%)**
* **USA GDP (Billions$ )-Quarterly Real GDP (adjusted for inflation)**
* **Unemployment (%)**
* **Delinquency Rate (%) on Mortgages (Foreclosure on the mortgage)-an indicator of the number of foreclosures in real estate.**

**EXPLORE THE DATASET**

* **Shape of supply dataset (Rows=250, columns=5)**
* **Shape of demand dataset (Rows=84, columns=6)**
* **Period 2005 the highest number of permits given to build, highest amount spent in the construction. Highest number of new housing projects started in the year 2005.**
* **There is highest number of sale in the year 2005.**
* **By univariate analysis**
* **After the year 2000 the increase of supply increases, till 2005.**
* **After the 2005 there is huge impact on supply, as we see it is drastically decreases almost 24% decrease in every year, till 2011. In 2011 comparison with 2005 almost 70 % decrease in supply.**
* **After the year 2011 there is again market will rise and almost 18% increase every year by the 2019.**
* **In the last 20 year 2020 is badly affected. In first quarter 60 % of supply reduced which is lowest in the last 20 year. So after the 2011 this is once in 2020 the supply is reduced.**
* **The impact of supply is clearly see in house price index.**
* **Correlation of supply data**

****

* **By the correlation the impact of one column clearly see in other. There is high correlation in supply data which means if number of the number of permits is increase then amount of spent for construction is increased , then new housing projects is also increased and then the number of selling of houses is also increases.**
* **One thing is clear from the supply data if one activity is increased or decreased it is directly impacted to other.**

**Explore the Demand data**

* **The unemployment rate is 13.2 % in the year 2020. Which is almost highest in the last 20 year. Previously in between 2009 to 2011 it is above 9 %.**
* **A Mortgage is a loan- provided by a mortgage lender or a bank that enables an individual to purchase a home or property. It is decreases per month.**
* **We see in univariate plot in 2000 it is above 8% which is highest in last 20 year. After the 2000 it is decreases in every month.**
* **2020 it is below 1 %.**
* **The growth of GDP increases every month. It is slightly decreases in 2010.**

**And in 2020 it is decrease by 12 percent in one quarter half. GDP is lowest in 2000 in last 20 year.**

* **Foreclosures is the action of taking possession of a mortgaged property when the mortgagor fails to keep up their mortgage payments. The process of transferring back the ownership of a home to the bank or lender after the borrower’s failure to repay the home loan.**
* **It is below 2% till 2007, in 2010 it is above 11 % which is highest in last 20 years. Between 2010 to 2013 foreclosures rate is above 10%. After 2013 again it is decreases which is good. Then in 2019 it is almost closed to 3% which almost closed to lowest in last 20 year. Again 2020 there is increases in rate.**
* **In 2020 HPI is 232.55 which is highest in last 20 year.**
* **In one year of 2020 there is increase of 10% in HPI.**
* **In 2012 HPI is 140 which is lowest in last 15 year.**

**Before perform the Machine Learning Model**

* **I have perform the feature engineering and data pre-processing then able to analysis that there is no null value in the data set. And some analysis done to detect the outliers but there is no outliers present in both the dataset.**
* **HPI is the target variable and HPI is normally distributed which means that data is good for perform machine learning model.**
* **Perform one – hot encoding.**
* **Perform Feature scaling**
* **Scaling the target value is a good idea in regression modelling, scaling of the makes it easy to for a model to learn and understand the problem.**

**Machine Learning Models**

**The primary purpose of machine learning models is to discover patterns in the user data and make predictions based on these and intricate patterns for answering business questions and solving business problems. Machine Learning helps in analyzing the data as well as identifying trends.**

**Split the data train and test. 85% data for training and 15 % for testing.**

**First Perform the Linear Regression Model**

* Regression searches for relationships among **variables**. For example, in this data observe HPI of USA and try to understand how the prices depend on their **factors**, such as unemployment rate, mortgage, GDP, and so on.
* This is a regression problem where data related to each HPI represents one **observation**. The presumption is that the unemployment rate, mortgage, GDP are the independent features, while the HPI depends on them.
* Typically, need regression to answer whether and how some phenomenon influences the other or how **several** variables are related. For example, use it to determine if and to what extent unemployment, GDP, mortgages impacts prices.
* **Linear Regression**
* Linear regression is probably one of the most important and widely used regression techniques. It’s among the simplest regression methods. One of its main advantages is the ease of interpreting results.

**Linear Regression Score**

* R square: 0.8295879940347317
* Adjusted R square: 0.8186641474984966
* MSE: 99.4449594433028
* RMSE: 9.972209356170918
* MAPE 0.04134175663523802

**Lasso Regression model score**

* R square: 0.8295665361137649
* Adjusted R square: 0.8186413140697755
* MSE: 99.45748134320183
* RMSE: 9.97283717621028
* MAPE 0.04134160464741266

Decision Tree Model scores

* R square: 0.9667979300890057
* Adjusted R square: 0.9646695922741984
* MSE: 19.375269230769252
* RMSE: 4.401734797868819
* MAPE 0.021496783563663583

Random Forest Model Score

* R square: 0.9343159395398215
* Adjusted R square: 0.9301054228436563
* MSE: 38.33033178346116
* RMSE: 6.191149471904322
* MAPE 0.017469097183405217

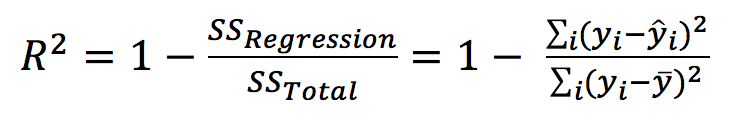
SVM Model score

* R square: 0.556345883425824
* Adjusted R square: 0.5279065169787615
* MSE: 258.8970500034206
* RMSE: 16.09027812076039
* MAPE 0.08023215554363122

Best metrics to evaluate Regression Model

## R Square, Adjusted R Square, MSE, RMSE, MAE

* Model evaluation is very important in data science. It helps to understand the performance of model and makes it easy to present your model to other people. There are many different evaluation metrics out there but only some of them are suitable to be used for regression
* **R Square/Adjusted R Square**
* R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.



* R Square is calculated by the sum of squared of prediction error divided by the total sum of the square which replaces the calculated prediction with mean. R Square value is between 0 to1 and a bigger value indicates a better fit between prediction and actual value.
* R Square is a good measure to determine how well the model fits the dependent variables. **However, it does not take into consideration of overfitting problem**. If your regression model has many independent variables, because the model is too complicated, it may fit very well to the training data but performs badly for testing data. That is why Adjusted R Square is introduced because it will penalize additional independent variables added to the model and adjust the metric to prevent overfitting issues.
* We can interpret that around 85% of dependent variability can be explained by the model, and adjusted R Square is roughly the same as R Square meaning the model is quite robust.

Conclusion

* **Out of all the model Decision Tree model gives the best result 96.3 %. Which means that this data set is good for decision tree model.**
* **All of the regression model gives the good result but the decision tree and random gives over 90%.**