Caitao Zhan, 111634527, caitao.zhan@stonybrook.edu

Report

Write a 2-3-page report about your favorite model. The report should include:

1. A description of how it works.
2. An evaluation of how well it works.
3. Any interesting experiences or surprises you had over the course of these experiments.

My favorit model is linear regression. Although it does not have the best result, it is the one I can best understand so far. Below is how it works.

Given a data set $\{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}, y_i\}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y_i and the vector of regressors x_{i1} is linear. What linear regression simply do is to compute the coefficients of $x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}$, to fit vector X to Y. So, when ever we are given a set of X, we can predict the value of Y.

I turn out that a simple linear regression works not bad! The logerror is 0.06507. This reflects the majic of KISS principle.

In linear regression models, we simply pick a subset of all features, and predict the objective logerror by assigning each feature a coefficient. First, we simply drop features that have too many missing values. The simple reason is that these features do not have enough data. We set a missing value threshold of 2,900,000.

After deleting the features with too many missing values, the number of features decreased from 58 to 41. The next step is to deal with catagorical features. A good way is recode categorical predictors to a binary numeric system. However, as a baseline model, it should aligns the KISS principle, so we decide to simply ingore them and delete them.

In the 28 features, there are some features that are highly correlated, sometimes they are even almost duplicated. Such kind of features do no good to our predition, so we decide to remove them and leave only one of the highly correlated features. After careful ovservation, the following are highly correlated features.

We need to make sure that there are no missing values in our final data. A simple imputation method is to replaces each missing value by the mean of the given column.In general, normally distributed values are the most comparable, so we also need to normalize the features. I tried Z-Score normalizing because this was what taught in class.

In file 'properties_2016_backup.csv', there are many properties of houses. However, there is no saling price. This is because saling price is business secret, and can not let others know. The target of prediction is instead logerror, which is in anohter file named 'train_2016_v2.csv'. So we skip the saling price and directly predict logerror. Since the predicting target and the features of houses are in two different spreadsheets, we need to join two spreadsheets together.

After all those wrangling with the data, finally we can do the linear regression! First, we need to separate the prediction target and the features. We need to make sure that the prediction target is logerror. For the feature part, do not forget to drop the parcelid. We do not want that during the regressions.

```
from from sklearn import linear_model

model = linear_model.LinearRegression()

model = model.fit(feature, target)
```

We are going to predict the same parcelid as in the original properties_2016.csv file, however, with a different date time. The date time are as follows: 2016/10, 2016/11, 2016/12, 2017/10, 2017/11, 2017/12. To do so, we first need to join the feature dataframe with it the sample_submission.csv, then add a new column named 'transactiondate', and fill it with the above six date time respectively. After the two dataframes are joined, we then drop the parcelid column. Predict six times: 2016/10, 2016/11, 2016/12, 2017/10, 2017/11, 2017/12. To preict the logerror of the month, a simple way is to predict the midth day of each month.

This first data science project is very rewarding! Before this project, I was even a Python rookie! After this project, I realize that I still have a long journey to go before becoming even a decent data scientist. Good news is that data science if fun, and interest drive a person to study harder and become a better person.

Reference:

[1]: https://en.wikipedia.org/wiki/Linear_regression