Machine learning project Dat158

Knut Erik Aspen, Group 24, 15.11.2022

DESCRIBE THE PROBLEM

SCOPE

The goal of this project is to go through a machine learning project from start to finish. I will be making and launching a machine learning model that can produce valuable results. For this project I chose to base myself on a dataset with house prices of the city Ames in Iowa from kaggle.com. In this project I will go through the “Housing Prices Competition for Kaggle Learn Users” competition and deploy my model.

This dataset is quite large and contains many different features. My plan is to train a model which can sort through and find the features which correlate most with the price of the houses. I plan on making a small webpage where a user can enter these values and get a reasonably accurate prediction from my model. For now I don't plan on including too many features on this web page, and only those deemed most important will be added. This can of course be expanded upon later.

Every business carries some inefficiencies that can be replaced by high-performing algorithms. For this project it seems reasonable to aim towards the real-estate business. Machine learning is already in full use in this field, and finding room for improvement was quite hard. For this reason I chose a rather simple scope which is very likely already in use today. I will aim at one of the jobs real-estate agents have, being home appraisal. With my machine learning model it should be possible to appraise your own home and get a reasonably accurate price prediction. This can save time and money for the real-estate business as they no longer would require as much manpower for appraising homes. Customers looking to sell their house could also be more willing to put their house on the market if getting a price estimate was as easy as pressing a few buttons.

For this project I will mainly use Python, kaggle notebook and many useful python libraries. Among them is Flask, which will be used for the deployment of my model.

METRICS

There are many metrics one can use to measure machine learning projects. The ones I will use to measure my system are largely based on the accuracy of my model. I believe as long as my model is reasonably accurate, it can already be used in the real-estate business to reduce costs and bring in more money compared to not having the option of “evaluating” your house online. I currently have no contact with any real-estate business, so metrics like revenue increase or customer interest increase will not be included even though they would be nice metrics to look at for this project. As I stated earlier I believe the accuracy of my model is most important and will use metrics to reflect that. In the kaggle workspace I tested many different models to find the best one for my dataset. I tested the mean squared error metric for all the models and eventually concluded Xgboost to be the best one for this project.

DATA

The data given in the kaggle competition has a training and a testing dataset. The training dataset consists of 1460 examples of houses with 81 features describing every house. The testing dataset consists of 1459 examples of houses with 79 features describing the houses. The testing set excludes the sale price because this is the value we are trying to predict in the competition. It also aligns nicely with my business goal to provide a useful model to the real-estate business.

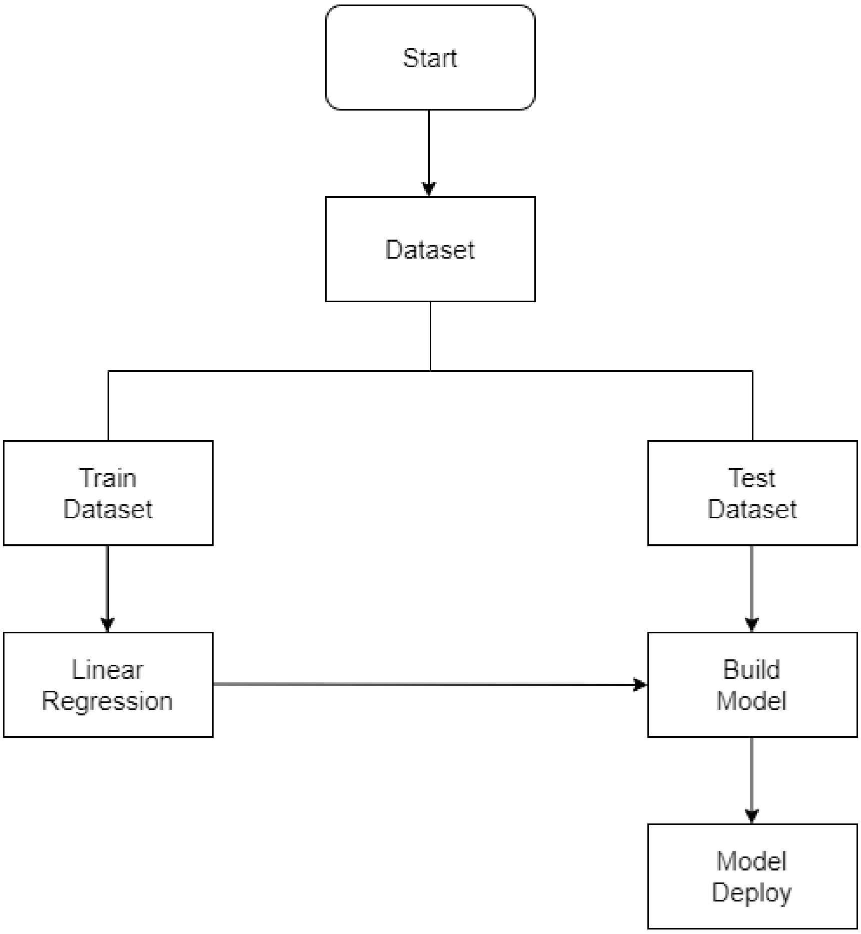
Machine learning tasks are usually split into three categories; supervised, unsupervised and reinforcement. For this project, the task is supervised learning. I have been given housing data consisting of features and labels, and I am tasked with predicting the labels for houses outside of the training data.

As any good machine learning project I started by exploring, then cleaning the data. The cleaning was done by checking for and removing features which were unneeded. Features with many null or missing values are safe to remove as they have little effect on predictions. The unneeded features were found by checking the relationships that exist within the data. I checked for the features with the highest correlation with the sale price. The values at the bottom of this list could safely be removed from the dataset without affecting predictions much if at all.

Machine learning models can't understand categorical data. I therefore needed to transform categorical data into numerical data. This was done by using one hot encoding.

MODELING

My task in this project is supervised learning. I know it is a regression task as I am tasked with predicting a numerical value. I therefore chose three machine learning models; Linear regression, random forest and gradient boosting. In short linear regression is analysis used to predict the value of a variable based on the value of another variable. Random forest uses multiple independent decision trees in parallel to learn from data and aggregates their predictions for an outcome. Gradient boosting is a type of boosting method that uses a combination of decision trees in series. Each tree is used to predict and correct the errors by the preceding tree additively. I will not use all these models, and rather test their mean squared error and choose the best one. In the end I chose gradient boosting for my model as this produced the best results.



DEPLOYMENT

The deployment of my model will be done mainly with Flask. I will make a small web page with a few fields for the user to enter. They will then get a prediction based on these fields. The scope of this web page will be quite small, but has a lot of room for expansion. The prediction will be more and more accurate with every field the user enters. The variables are currently set to base values instead of the user having to enter every single feature still included within the dataset. As I stated earlier only the 4 variables with highest correlation with the sale price are currently implemented.

REFERENCES

Model references:
https://www.kaggle.com/competitions/home-data-for-ml-course/overview
https://www.kaggle.com/code/sumeetmalusare/first-ml-project-house-prediction-using-xg-boost

Raport references:

https://towardsdatascience.com/predicting-house-prices-with-machine-learning-62d5bcd0d68f

Python/Flask references:
https://medium.com/analytics-vidhya/deploy-house-price-prediction-using-flask-16d88c40d0cd
https://github.com/alu042/DAT158-2022/tree/main/a_quick_flask_tutorial