

Capstone Proposal – Become a ML Engineer:

Domain background: The field of research where the project is derived is real estate, with the ambition to successfully predict the value of houses in Ames (Iowa)¹.

Problem stating: Numerous combinations of factors affect the price of a house in Ames. Given this unclarity this problem will be solved by finding out which factors are key in determining the price of a house and develop a tailor-made tool able to predict the potential price of a desired house determined by customizable characteristics.

Datasets and inputs:

Kaggle competition: House Prices: Advanced Regression Techniques -
(<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview>)

Dataset contains 80 variables directly related to property sales with qualitative and quantitative indicators of many physical attributes of the property.

Aim: predict value of SalePrice variable.

Solution statement: Create a web application that will have a front-end giving the user the ability to customize his desired house characteristics by providing relevant information (user-input) for the top-10 variables for the House Prices dataset.

This information will be sent through an API to a previously trained SageMaker pytorch Neural Network that will make the prediction and send back to the front-end the estimated price of that house.

Benchmark model: Kaggle's leaderboard and kernels offer plenty of resources to whom compare my results against. Making a submission to Kaggle will show the goodness of my model.

Evaluation metrics: MSE – mean squared error, comparing true values of houses in the test set versus predictions for houses in that same test set.

Project design:

- Sagemaker Jupyter notebook to:
 - Ingest and clean raw data.
 - Perform feature selection algorithm (Univariate selection)
 - Uploading data to S3
 - Training a pytorch Neural Network
 - Evaluating goodness of the model (MSE)
 - Deploy model
- Model.py, train.py, predict.py storing all the necessary lines of code to construct the architecture of the NN, train it and make predictions
- Lambda_function.py file to upload to AWS
- API Gateway to create API for frontend and lambda to communicate
- Index.html for front-end/visual purposes

¹ <http://jse.amstat.org/v19n3/decock.pdf>