Project Proposal

SP18: APPLIED MACHINE LEARNING: 33910

Submitted By: Murali Cheruvu

Project title: Predicting Housing Prices

Project abstract in 150 words (remember to use the STAR methodology)

Situation: Predicting Housing Sale Prices using publicly available training and test datasets

Task: Apply exploratory analysis and various machine learning algorithms

Action: Build ML pipeline to achieve exploratory analysis and run various ML algorithms.

Result: Predict sale prices of the houses in the test dataset and measure accuracy of the outcome

Data you plan to use (number and type of inputs and outputs, size of data, data publicly available or not; how was the data acquired and by whom; do you have a data dictionary)

Kaggle Website provides the dataset for the housing prices. In this project, I would like to predict sale prices of housing prices using two datasets - training and testing, each with 79 exploratory variables describing almost every aspect of residential homes in city of Ames, Iowa State. However, these datasets are snapshots taken in 2010. As a result, these datasets may not reflect the latest trends in the housing sale prices but the analytical approaches taken in this project are generic and can easily be applied to newer datasets.

More details can be found here: https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Can this be a group project where 2-3 of your classmates can join you?

I would like to do this project on my own.

Which machine learning algorithm are you considering to use and why?

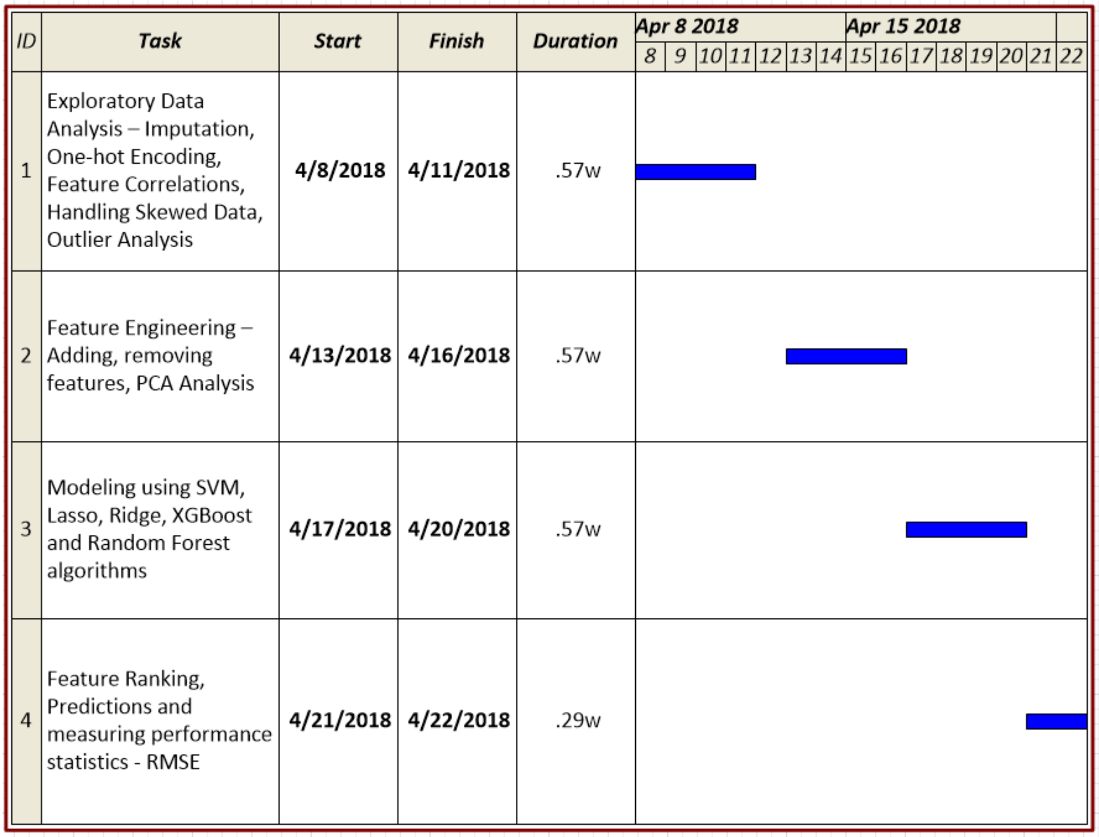
Predicting Housing Sale Prices is a Regression Problem. I would like to apply various machine learning algorithms including Lasso, Ridge, XGBoost and Random Forest algorithms using scikit-learn Python ML libraries.

Lots of participants of this Kaggle Competition have tried these algorithms to achieve optimized model/processing. Ultimately model ensembling gives better performance of the unified predictions.

Can you describe the metrics that you might use to measure success (standard metrics and domain specific metrics)?

I would like to measure the predictions using Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sale price.

Please provide a block diagram (Gantt diagram) of the key steps involved in completing this task and timeline



ID	Task	Start	Finish	Duration	*Apr 8 2018*							*Apr 15 2018*							
					8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	Exploratory Data Analysis – Imputation, One-hot Encoding, Feature Correlations, Handling Skewed Data, Outlier Analysis	**4/8/2018**	**4/11/2018**	.57w	■	■	■	■											
2	Feature Engineering – Adding, removing features, PCA Analysis	**4/13/2018**	**4/16/2018**	.57w						■	■	■	■						
3	Modeling using SVM, Lasso, Ridge, XGBoost and Random Forest algorithms	**4/17/2018**	**4/20/2018**	.57w										■	■	■	■		
4	Feature Ranking, Predictions and measuring performance statistics - RMSE	**4/21/2018**	**4/22/2018**	.29w														■	■

Provide a description of the machine learning pipeline that you plan to use

I am planning to use (1) various exploratory analysis activities including: imputation of null data, correlations of features, one-hot encoding to convert categorical features to numerical, analyze and fix skewed data and perform outlier analysis, and then (2) feature engineering: to add/remove features to enrich the model. Perform PCA analysis to identify key features drive most of the variations and highly correlated with target variable – sale price. After that (3) apply various machine learning algorithms; tune hyper-parameters of these algorithms and measure RMSE of the predictions for accuracy.