House Price Advanced Prediction

Hello!

Zia Khan

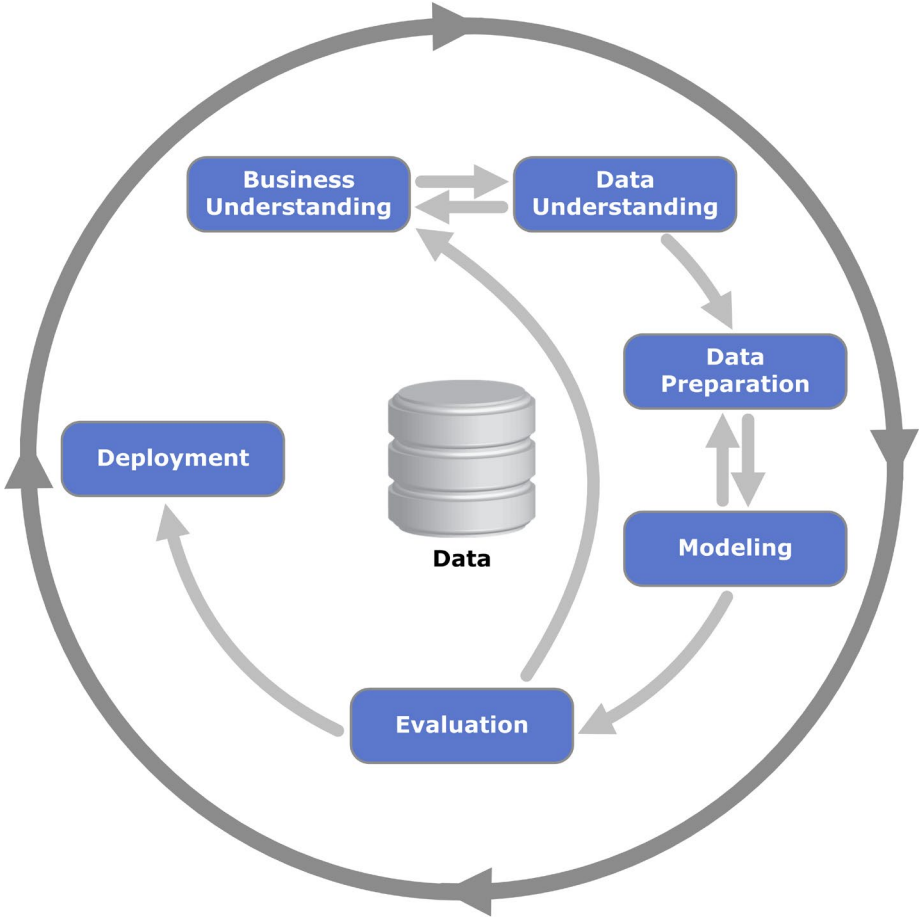
zia@thedevmasters.com

CRISP-DM	1
Business Objective	2
Data Understanding	3
Data Preparation	4
Modeling	5

CRISP-DM

Overview

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment



Kaggle Competitions

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

CRISP-DM 1

Business Objective 2

Data Understanding 3

Data Preparation 4

Modeling 5

2. Objective

Predict the price of a house based on the dataset from Kaggle

CRISP-DM 1

Business Objective 2

Data Understanding 3

Data Preparation 4

Modeling 5

Data Understanding

Data Introduction

File descriptions

- **train.csv** - the training set
- **test.csv** - the test set
- **data_description.txt** - full description of each column, originally prepared by Dean De Cock but lightly edited to match the column names used here
- **sample_submission.csv** - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

Data Understanding

```
df_train.columns
```

```
Index([u'Id', u'MSSubClass', u'MSZoning', u'LotFrontage', u'LotArea',
       u'Street', u'Alley', u'LotShape', u'LandContour', u'Utilities',
       u'LotConfig', u'LandSlope', u'Neighborhood', u'Condition1',
       u'Condition2', u'BldgType', u'HouseStyle', u'OverallQual',
       u'OverallCond', u'YearBuilt', u'YearRemodAdd', u'RoofStyle',
       u'RoofMatl', u'Exterior1st', u'Exterior2nd', u'MasVnrType',
       u'MasVnrArea', u'ExterQual', u'ExterCond', u'Foundation', u'BsmtQual',
       u'BsmtCond', u'BsmtExposure', u'BsmtFinType1', u'BsmtFinSF1',
       u'BsmtFinType2', u'BsmtFinSF2', u'BsmtUnfSF', u'TotalBsmtSF',
       u'Heating', u'HeatingQC', u'CentralAir', u'Electrical', u'1stFlrSF',
       u'2ndFlrSF', u'LowQualFinSF', u'GrLivArea', u'BsmtFullBath',
       u'BsmtHalfBath', u'FullBath', u'HalfBath', u'BedroomAbvGr',
       u'KitchenAbvGr', u'KitchenQual', u'TotRmsAbvGrd', u'Functional',
       u'Fireplaces', u'FireplaceQu', u'GarageType', u'GarageYrBlt',
       u'GarageFinish', u'GarageCars', u'GarageArea', u'GarageQual',
       u'GarageCond', u'PavedDrive', u'WoodDeckSF', u'OpenPorchSF',
       u'EnclosedPorch', u'3SsnPorch', u'ScreenPorch', u'PoolArea', u'PoolQC',
       u'Fence', u'MiscFeature', u'MiscVal', u'MoSold', u'YrSold', u'SaleType',
       u'SaleCondition', u'SalePrice'],
      dtype='object')
```

CRISP-DM	1
Business Objective	2
Data Understanding	3
Data Preparation	**4**
Modeling	5

Data Preparation

- Missing values
- Categorical variables
- Numeric variables
- Ranking variables
- Feature engineering
- Correlation
- Selecting the best variable

CRISP-DM	1
Business Objective	2
Data Understanding	3
Data Preparation	4
Modeling	**5**

Modeling

- Combine data (train and test)
- Clean and fill missing values
- Separate train and test dataset
- Split the train dataset into train and test
- Train model and test
- Use Gradient Boosting and Random forest regressors
- Mean absolute error or mean square error
- Predict on actual Kaggle test dataset

Thanks!!

Any questions?

zia@thedevmasters.com