

A pair of black-rimmed glasses is resting on a stack of books. A red bookmark is visible between the pages of the books. The background is slightly blurred, emphasizing the glasses and the text.

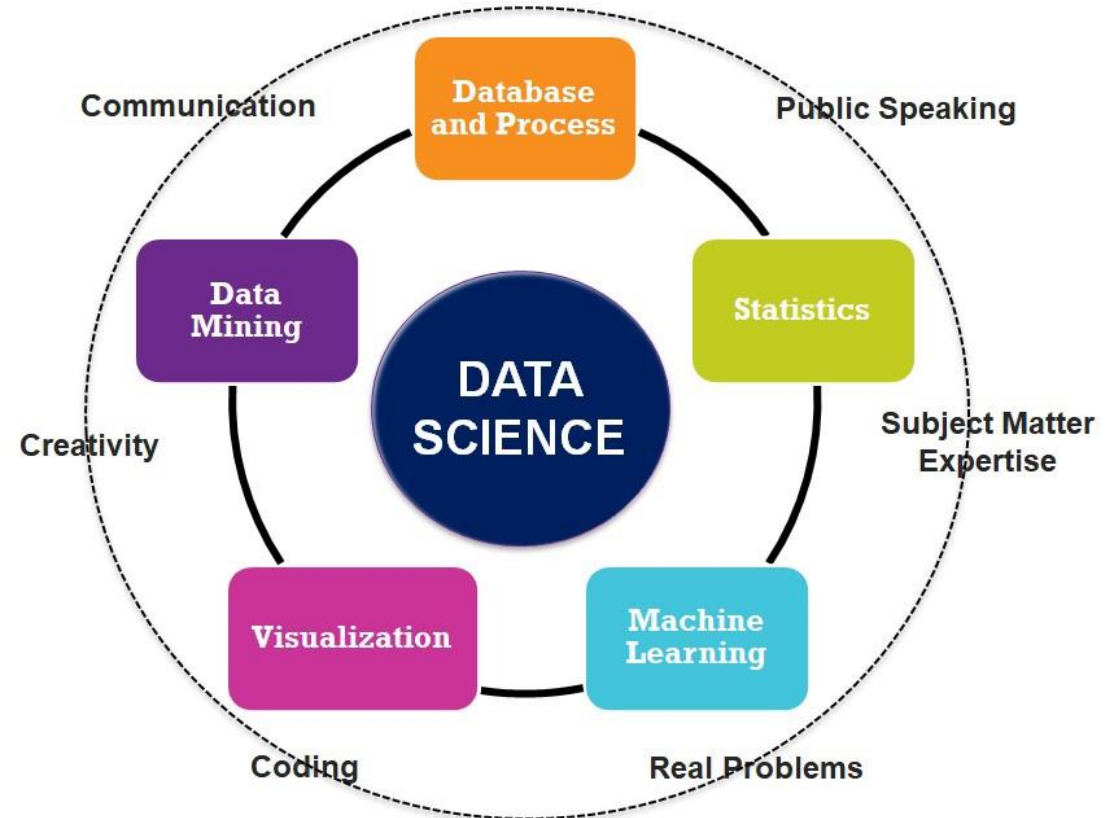
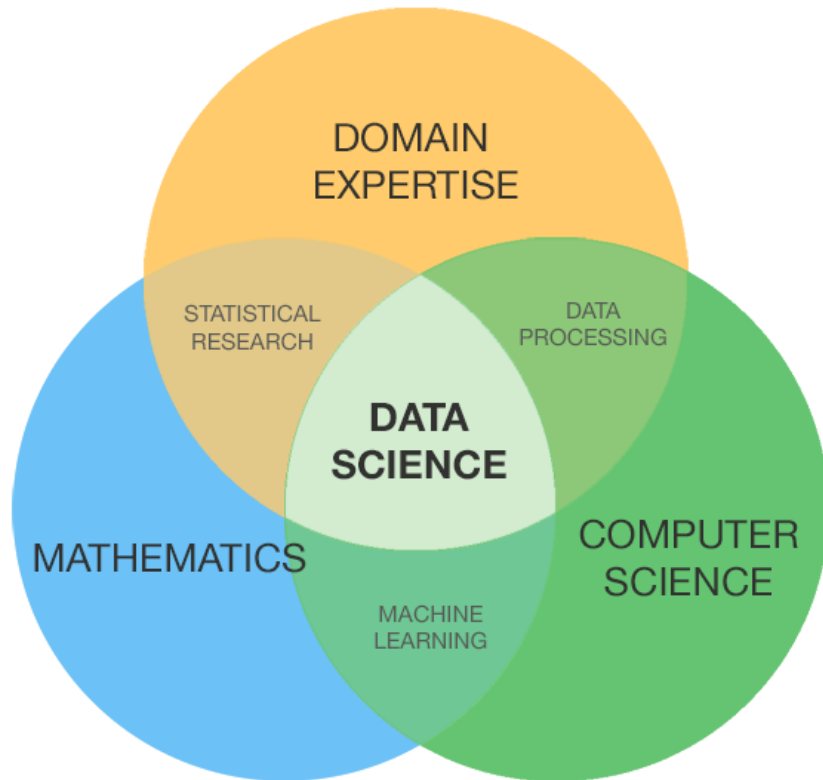
Data Science

Class 1: Introduction to Data Science

US – Embassy

Preface

- "There are many skills under the umbrella of data science, and we should not expect any one single person to be a master of them all".*



Vision & missions of Data science

Programming
language



Vision & Mission.

- 1) Viewing & analyzing dataset
- 2) Cleaning, transforming and extracting dataset
- 3) Model selection
- 4) Understanding metrics and scores in valuating model.
- 5) Detect overfitting and how to deal with
- 6) Understanding NLP (Natural Language Processing), Computer vision, Image Processing, also some Machine Learning algorithms and Deep Learning model via CNN (convolution neural network).

Problems

Problems / examples.

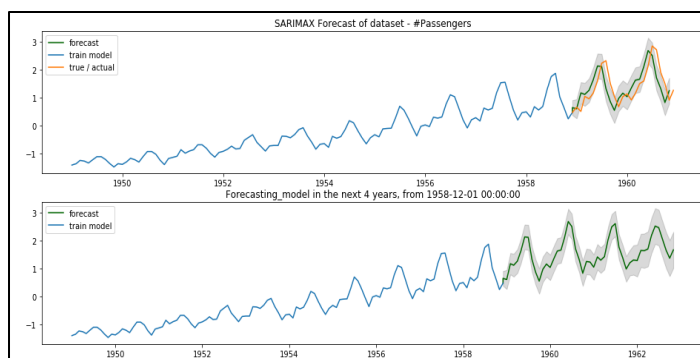
- 1) Spam message detection.
- 2) Dog (cat) breed prediction; Human face detection.
- 3) Making recommendations.
- 4) Consumer behavior segmentation
- 5) Demand forecasting
- 6) Telecom / Bank customer churn prediction
- 7) GANs (Generative Adversarial Networks)
- 8) Gaming World
- etc

Type of issues

- 1) Classification
- 2) Segmentation
- 3) Time series analysis
- 4) Sentiment analysis
- 5) Dimension reduction
- 6) Regression
- etc.



Fake Images ((lr_D = 0.000180, lr_G = 0.000220; 200 epochs, 32 batch-size)



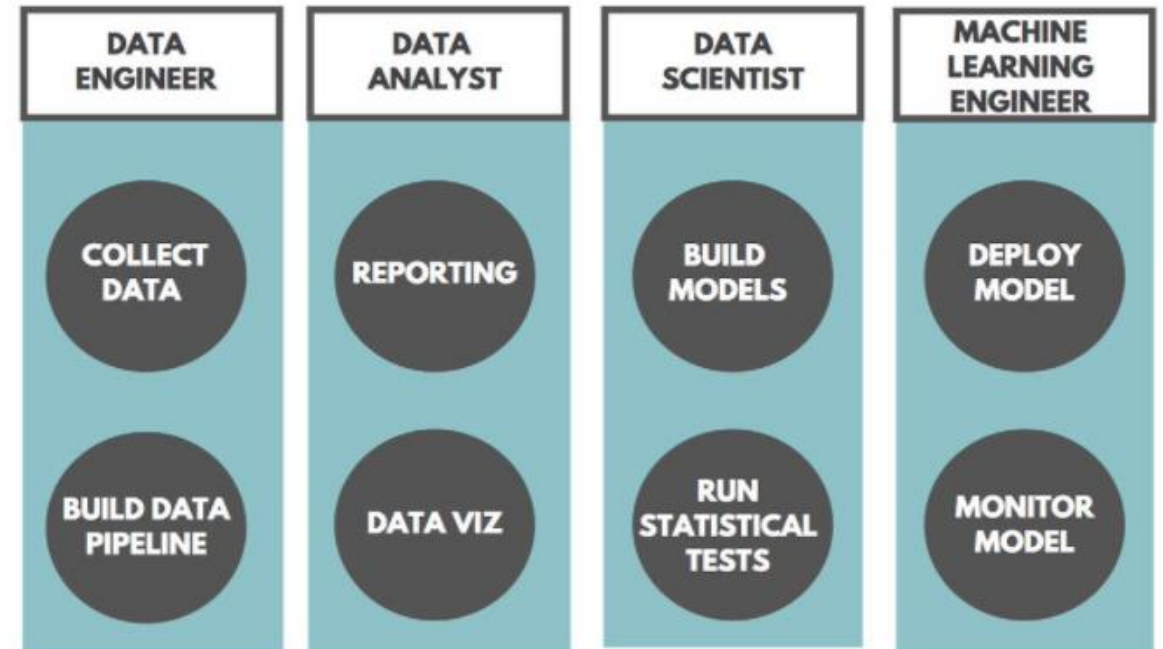
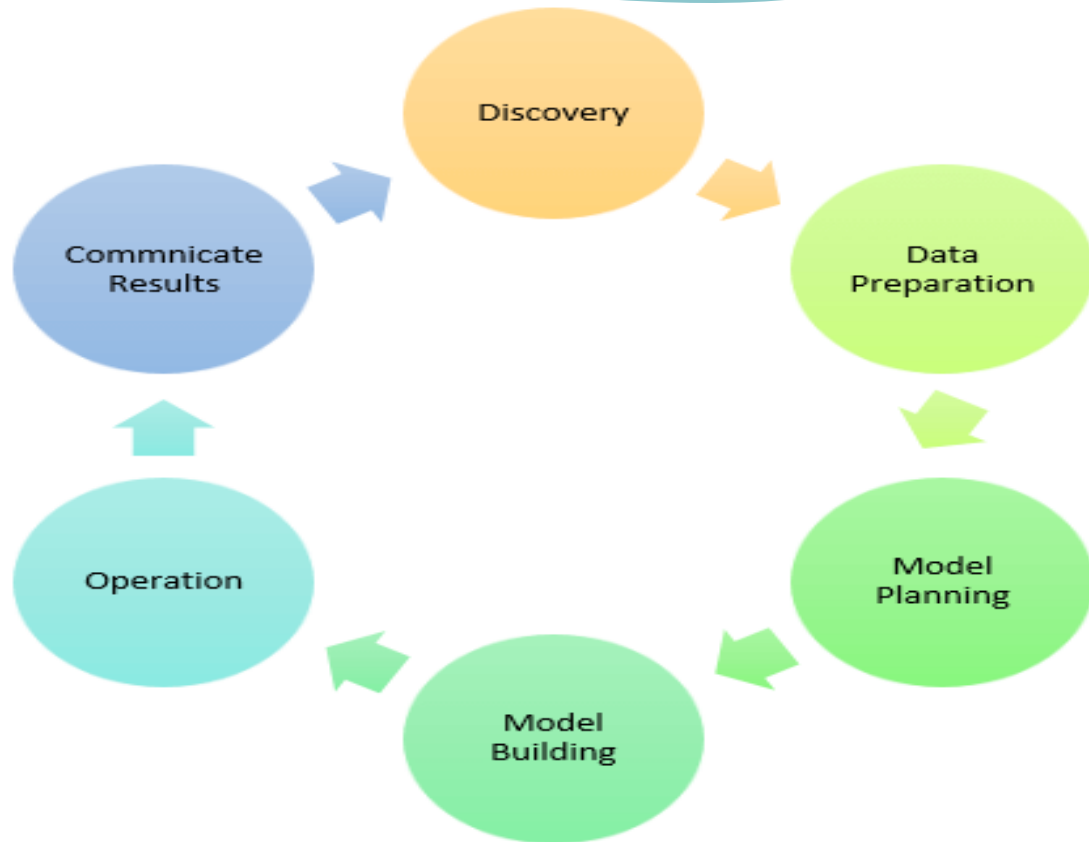
```
image_path = 'images/' + image_list[8]  
predict_breed(image_path)
```



Dog and human were detected in this image
If this one were a dog, he / she would be a ... Golden_retriever!!

Data Science process

- Exploring dataset,
- Data Preparation,
- Model Planning,
- Model Building,
- Operationalize,
- Communicate Results



Challenges

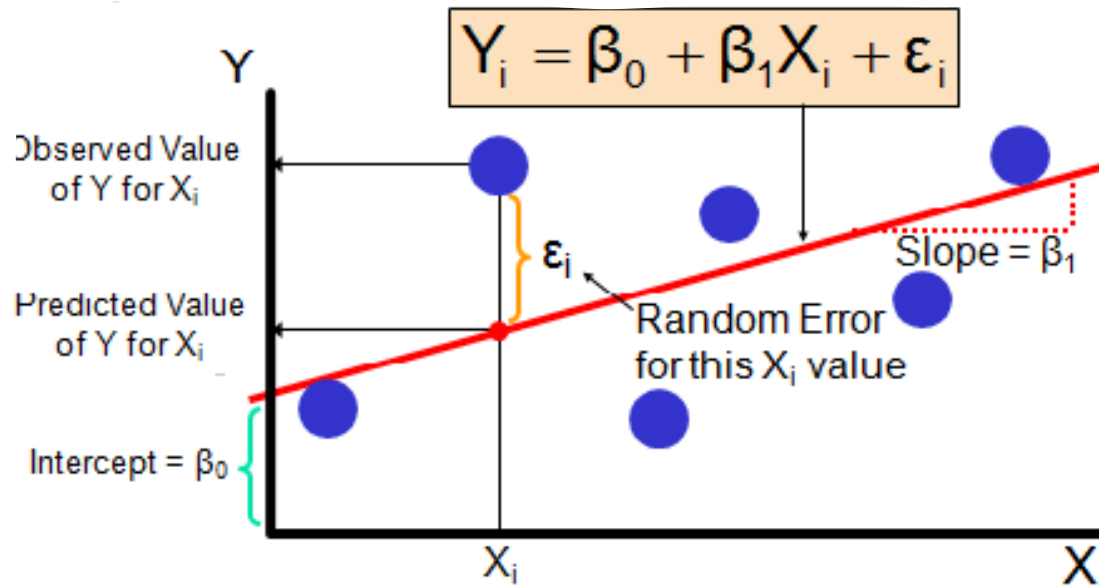
- High variety of information & data is required for accurate analysis
- Not adequate data science talent pool available
- Management does not provide financial support for a data science team
- Unavailability of/difficult access to data
- Data Science results not effectively used by business decision makers
- Explaining data science to others is difficult
- Privacy issues
- Lack of significant domain expert
- If an organization is very small, they can't have a Data Science team.



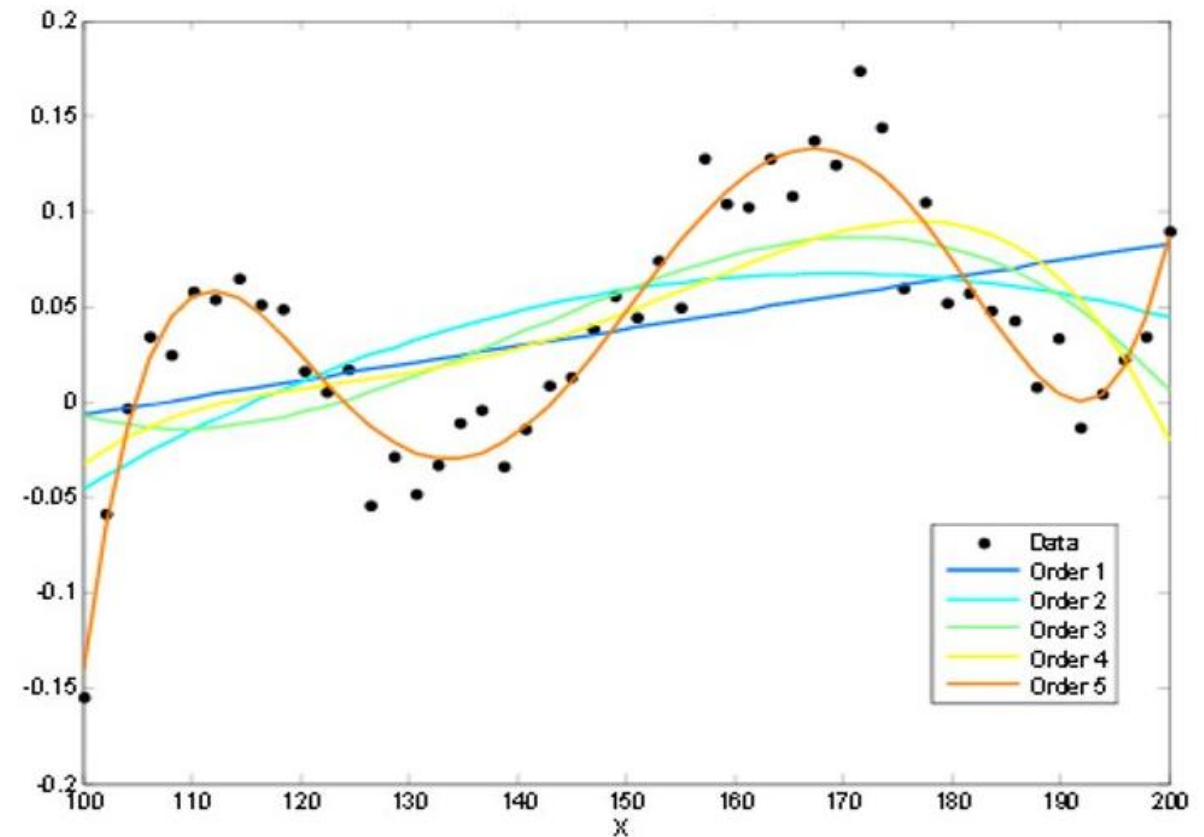
Recap.

- Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes.
- Statistics, Visualization, Deep Learning, Machine Learning, are important Data Science concepts.
- Data Science Process goes through Discovery, Data Preparation, Model Planning, Model Building, Operationalize, Communicate Results.
- Important Data Scientist job roles are: Data Scientist, Data Engineer, Data Analyst, Statistician, Data Architect, Business Analyst, etc.
- Python, R, Sparks, SQL, SaS, are essential Data science tools.
- The predictions of Business Intelligence is looking backward while for Data Science it is looking forward.
- Important applications of Data science are : Image & Speech Recognition, Recommendation Systems, Demand forecasting, Gaming world, Online Price Comparisons, Churn prediction, etc.
- High variety of information & data is the biggest challenge of Data Science technology.

Regression line



Simple Linear Regression



Simple Polynomial regression

Regression model

Definition.

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as the independent variables).

$$Y = f(X_1, X_2, \dots, X_n)$$

Classify regression problem

- - Simple linear regression :

$$Y = a_0 + a_1 X_1$$

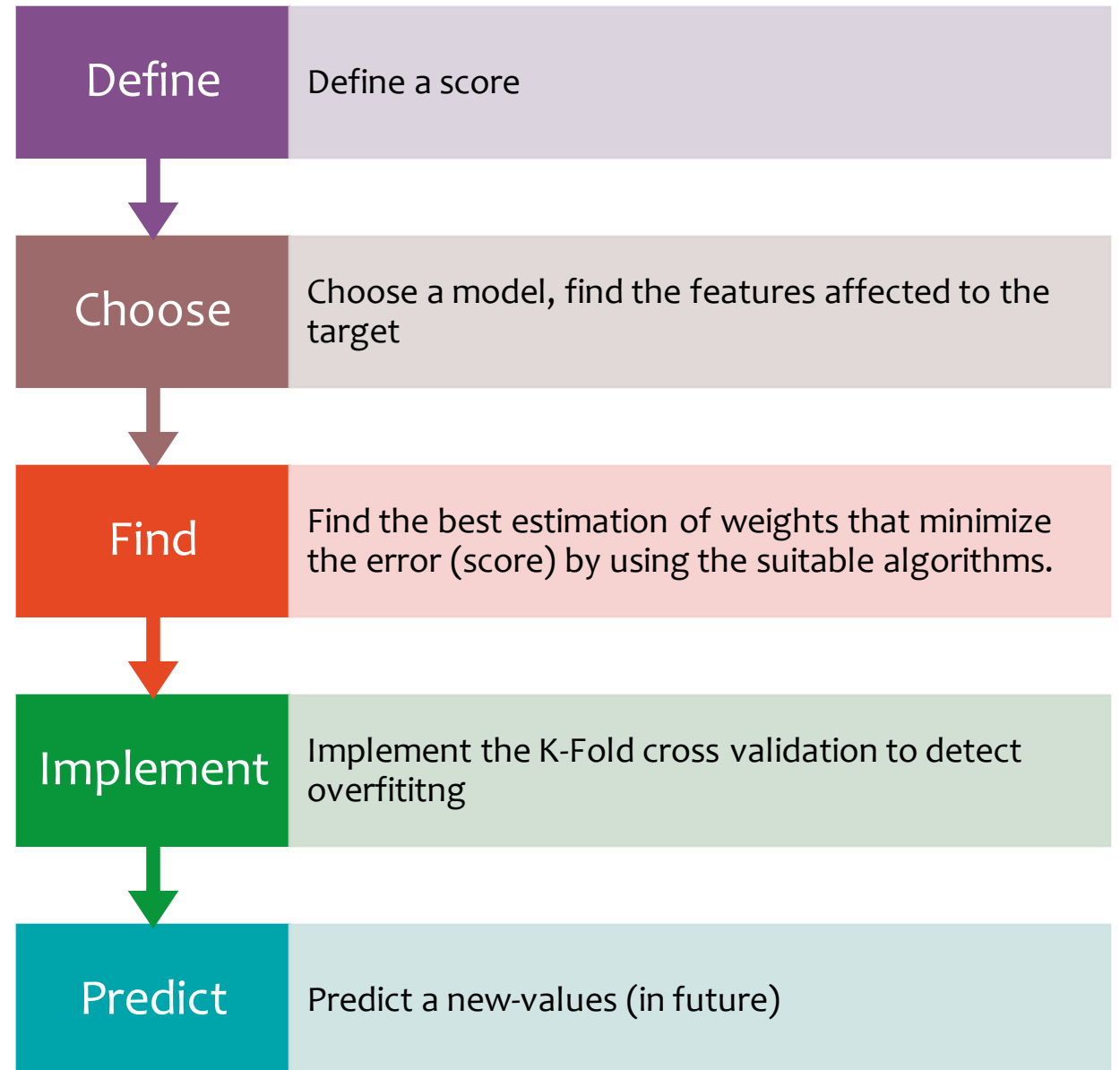
- - Multiple linear regression :

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \dots + a_n X_n$$

- - Non linear regression :

$$Y = f(X_1, X_2, \dots, X_n); \text{ e.g } Y = 2 - 3\log(X_1 + X_2) + 2 X_1 X_2$$

Regression model's objectives.



Regression algorithms in ML.

Regression function

Linear Regression

Ridge regression

Lasso regression

RandomForestRegressor

ExtraTree Regressor

Gradient Boosting
Regressor

Ensemble approaches

AdaBoost Regressor

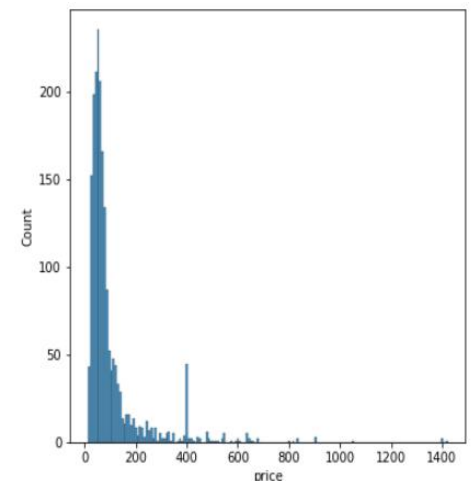
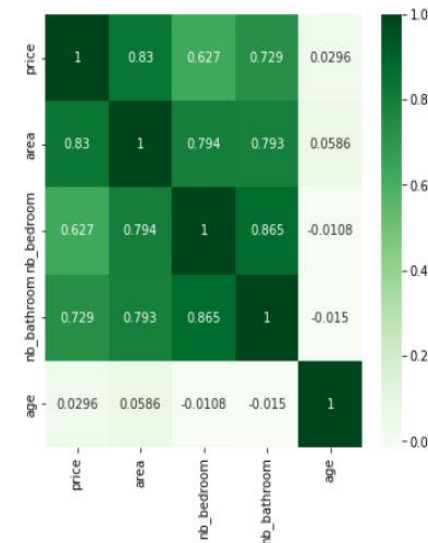
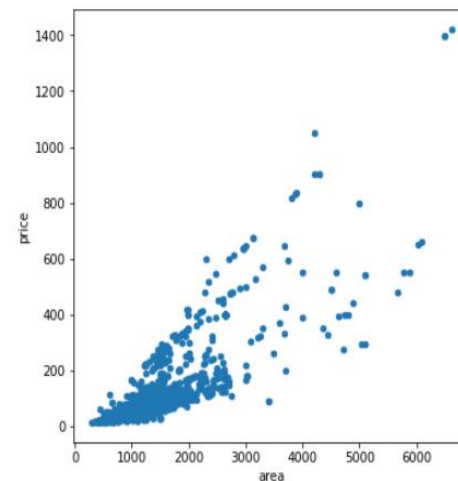
etc.

EDA technique

- This step helps you find out which features be highly correlated to the target.
- Depends on which assumptions in your model, such as linear regression
 - **Linearity:** The relationship between X and the mean of Y is linear.
 - **Homoscedasticity:** The variance of residual is the same for any value of X
 - **Independence:** Observations are independent of each other.
 - **Normality:** For any fixed value of X; the target Y is normally distributed.

In the regression problem, you can looking for the

- Scatter-plot,
- Heat-map correlation matrix,
- Histogram, density plot.



Metrics evaluation

The higher the better, 1 the best.

- R^2_{score} , $R^2_{\text{score_adj}}$
- Pearson correlation coef

The lower the better.

- MAE vs MAPE
- AIC, BIC
- MSE, RMSE or RLMSE

$$\begin{aligned} R^2_{\text{score}} \text{ (coefficient of determinant)} &= 1 - \frac{SSR}{SST} \\ R^2_{\text{adj}} \text{ (Adjusted R Squared)} &= 1 - \left(\frac{(n-1)(1-R^2_{\text{score}})}{n-k-1} \right) \\ \text{MAPE (Mean Absolute Percentage Error)} &= \frac{1}{n} \sum_{k=1}^n \left| \frac{y_k^{\text{true}} - y_k^{\text{pred}}}{y_k^{\text{true}}} \right| \\ \text{MAE (Mean Absolute Error)} &= \frac{1}{n} \sum_{k=1}^n |y_k^{\text{true}} - y_k^{\text{pred}}| \\ \text{MSE (Mean Squared Error)} &= \frac{1}{n} \sum_{k=1}^n (y_k^{\text{true}} - y_k^{\text{pred}})^2 \\ \text{RMSE (Root Mean Squared Error)} &= \sqrt{\text{MSE}} \\ \text{RMSLE (Root Mean Squared Logarithmic Error)} &= \frac{1}{n} \sqrt{\sum_{k=1}^n (\log(y_k^{\text{pred}} + 1) - \log(y_k^{\text{true}} + 1))^2} \\ \text{AIC (Akaike Information Criterion)} &= \frac{2}{n} (k - LL) \\ \text{BIC (Bayesian Information Criterion)} &= -2LL + k \log(n) \end{aligned}$$

- n is the number of observations,
- k is the number of parameters in the model.
- LL is the log-likelihood of the model
- y_k^{true} is the actual values k^{th} , and y_k^{pred} is the forecasted values k^{th} from y_k^{true}
- SSR meant **sum of squares total**, is the squared differences between

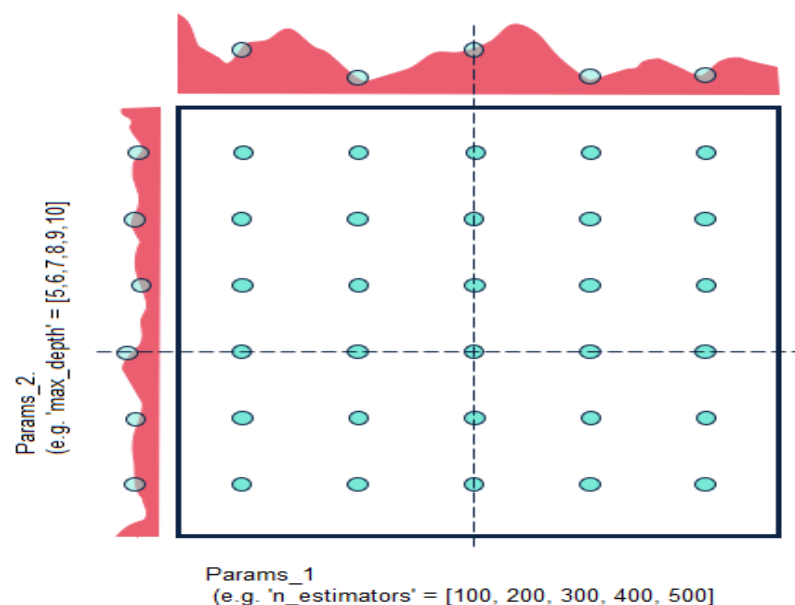
the observed dependent variable and its mean.

$$SST = \frac{1}{n} \sum_{k=1}^n (y_k^{\text{true}} - \overline{y^{\text{true}}})^2$$

- SSR is the **sum of squares due to regression**,

$$SSR = \frac{1}{n} \sum_{k=1}^n (y_k^{\text{pred}} - \overline{y^{\text{true}}})^2$$

Grid-search CV & Random-search CV

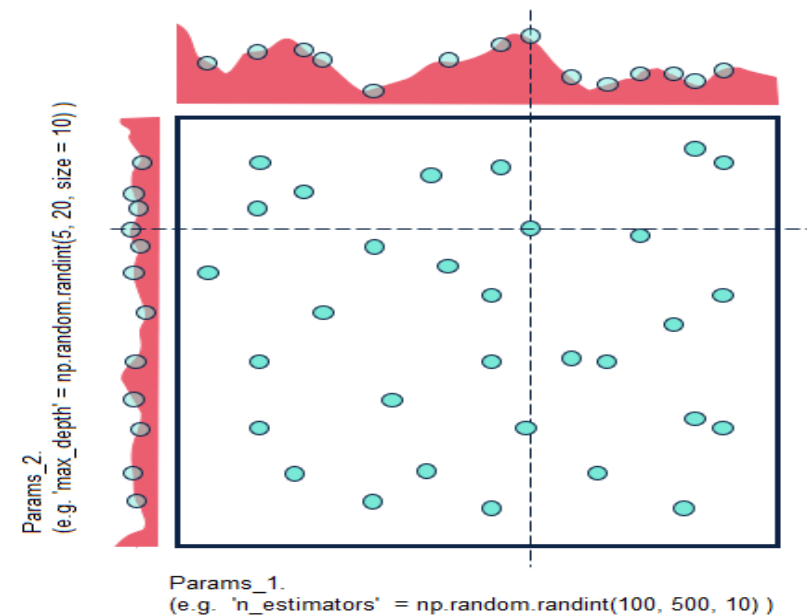


Grid Search

```
from sklearn.model_selection import GridSearchCV, RandomizedSearchCV
from sklearn.ensemble import RandomForestRegressor

grid_params = {
    'n_estimators': [100, 200, 300, 400, 500],
    'max_depth': [5, 6, 7, 8, 9, 10]
}
clf = RandomForestRegressor()
GridSearchCV(clf, param_grid = grid_params)
```

```
GridSearchCV(estimator=RandomForestRegressor(),
              param_grid={'max_depth': [5, 6, 7, 8, 9, 10],
                           'n_estimators': [100, 200, 300, 400, 500]})
```



Random Search

```
from sklearn.model_selection import GridSearchCV, RandomizedSearchCV
from sklearn.ensemble import RandomForestRegressor

grid_params = {
    'n_estimators': np.random.randint(100, 500, 10),
    'max_depth': np.random.randint(5, 20, 10)
}
clf = RandomForestRegressor()
RandomizedSearchCV(clf, param_distributions = grid_params)
```

```
RandomizedSearchCV(estimator=RandomForestRegressor(),
                    param_distributions={'max_depth': array([13, 8, 6, 7, 15, 18, 10,
                                                             'n_estimators': array([445, 222, 252, 287, 116,
```

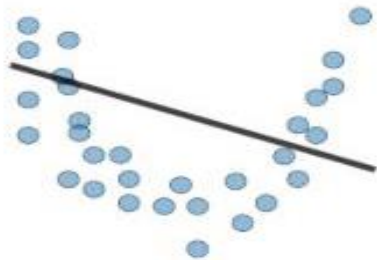


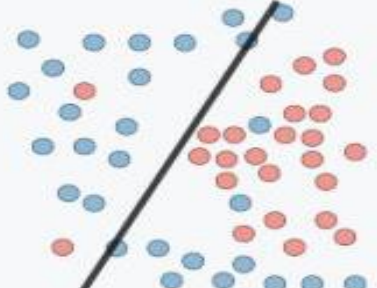
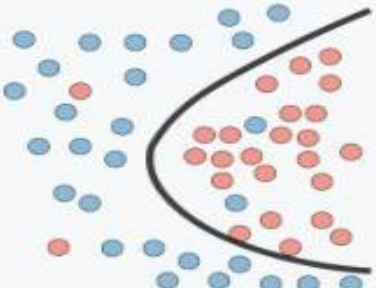
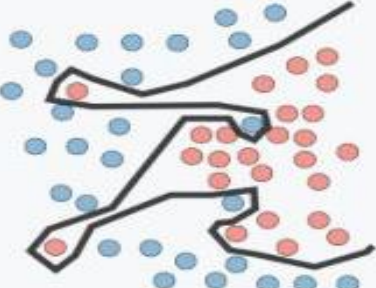


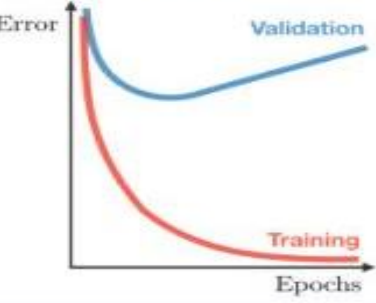
Grid-search CV & Random-search CV

You can add K-Fold cross validate into the Grid-Search CV to evaluate the standard deviation of score on the k folds. So,

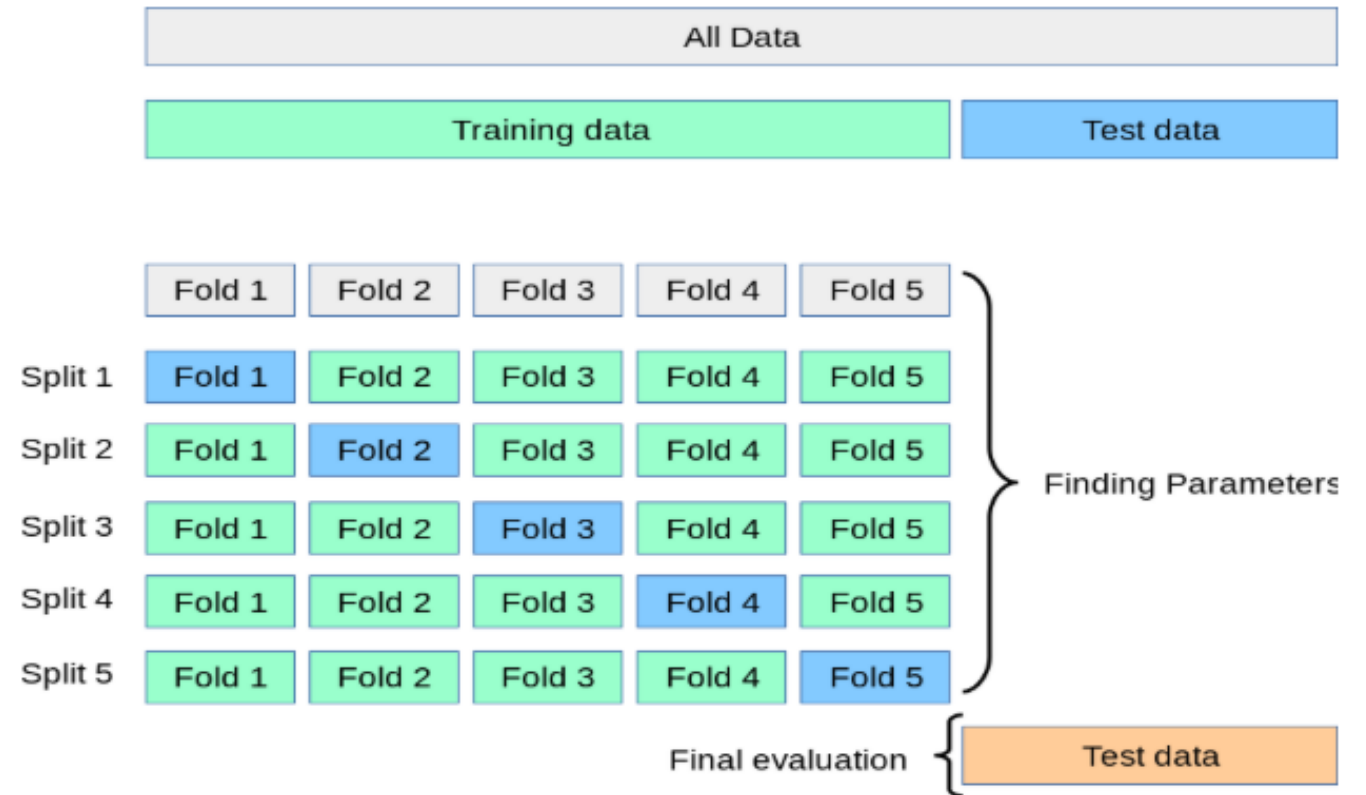
- What is overfitting, underfitting?
- What is K-folds croos-validation?

	MLA_name	train-test.shape	trained_time	best_params	std_score_kfolds	train_score	test_score	MAPE_test	MAE_test	MSE_test	Pearson_corr
8	GradientBoostingRegressor	((1557, 5), (566, 5))	73 mins 31.74 seconds	{'learning_rate': 0.1, 'loss': 'absolute_error...	0.023069	0.801824	0.756501	0.048558	0.089539	0.016794	0.871646
6	ExtraTreesRegressor	((1557, 5), (566, 5))	0 mins 28.37 seconds	{'bootstrap': True, 'criterion': 'squared_erro...	0.029491	0.816491	0.753660	0.053762	0.098010	0.016990	0.870965
3	ExtraTreesRegressor	((1557, 3), (566, 3))	1 mins 9.84 seconds	{'bootstrap': True, 'criterion': 'squared_erro...	0.036284	0.848059	0.753334	0.053052	0.096948	0.017012	0.870727
2	Lasso	((1557, 3), (566, 3))	0 mins 1.61 seconds	{'alpha': 0.01, 'max_iter': 100, 'selection': ...	0.014519	0.742205	0.749616	0.054238	0.098362	0.017269	0.867035
7	LinearRegression	((1557, 5), (566, 5))	0 mins 0.3 seconds	{'n_jobs': 1, 'normalize': True}	0.021541	0.747008	0.745725	0.054598	0.099642	0.017537	0.868468
1	Ridge	((1557, 3), (566, 3))	0 mins 4.01 seconds	{'alpha': 0.1, 'max_iter': 100, 'solver': 'lsq...	0.021199	0.747025	0.745715	0.054672	0.099710	0.017538	0.868620
0	LinearRegression	((1557, 3), (566, 3))	0 mins 0.27 seconds	{'n_jobs': 1, 'normalize': False}	0.021413	0.747025	0.745603	0.054694	0.099749	0.017546	0.868617
5	RandomForestRegressor	((1557, 5), (566, 5))	0 mins 49.28 seconds	{'bootstrap': True, 'max_depth': 7, 'min_sampl...	0.029588	0.841202	0.744372	0.053724	0.098667	0.017631	0.869372
4	RandomForestRegressor	((1557, 3), (566, 3))	0 mins 45.67 seconds	{'bootstrap': True, 'max_depth': 7, 'min_sampl...	0.033120	0.838689	0.742502	0.053797	0.098867	0.017759	0.867713
9	AdaBoostRegressor	((1557, 5), (566, 5))	0 mins 17.15 seconds	{'learning_rate': 0.001, 'n_estimators': 200}	0.021376	0.753175	0.715913	0.058387	0.105483	0.019593	0.854567

Overfitting and underfitting.

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none">• High training error• Training error close to test error• High bias	<ul style="list-style-type: none">• Training error slightly lower than test error	<ul style="list-style-type: none">• Very low training error• Training error much lower than test error• High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	<ul style="list-style-type: none">• Complexify model• Add more features• Train longer		<ul style="list-style-type: none">• Perform regularization• Get more data

How to prevent overfitting



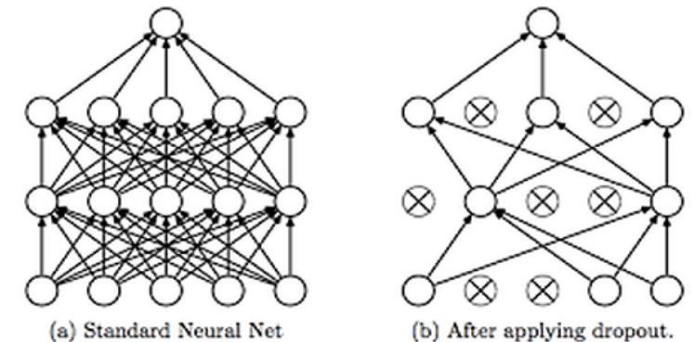
Validation: validation, cross-validation, Leave one out, KFold cross validation, etc

Data simplification : DropOut, weight decay

Training with more data

Regularization

Early Stopping



Regularization

$$L(x_1, \dots, x_n, y) = \sum_{k=1}^n (y_k^{\text{true}} - F_{\beta}^{\text{estimate}}(x_1, \dots, x_n))^2 + \lambda \sum_{k=1}^n |\beta_k|$$

$$L(x_1, \dots, x_n, y) = \sum_{k=1}^n (y_k^{\text{true}} - F_{\beta}^{\text{estimate}}(x_1, \dots, x_n))^2 + \lambda \sum_{k=1}^n \beta_k^2$$

L1 Regularization	L2 Regularization
1. L1 penalizes sum of absolute values of weights.	1. L2 penalizes sum of square values of weights.
2. L1 generates model that is simple and interpretable.	2. L2 regularization is able to learn complex data patterns.
3. L1 is robust to outliers.	3. L2 is not robust to outliers.

Summary

—

THE END

Thank for your considerations.

Data Science 's mission, process and challenges

Regression model and evaluated-scores

Algorithms used in Regression problems

Grid-Search CV and Random-Search CV

Definition of overfitting, underfitting

How to prevent overfitting

K-Folds cross validate and Regularization.

References

- [Metric evaluation](#)
- [Linear model](#)
- [Ensembling method](#)
- [Ensemble approach in regression](#)
- <https://alex.smola.org/drafts/thebook.pdf>