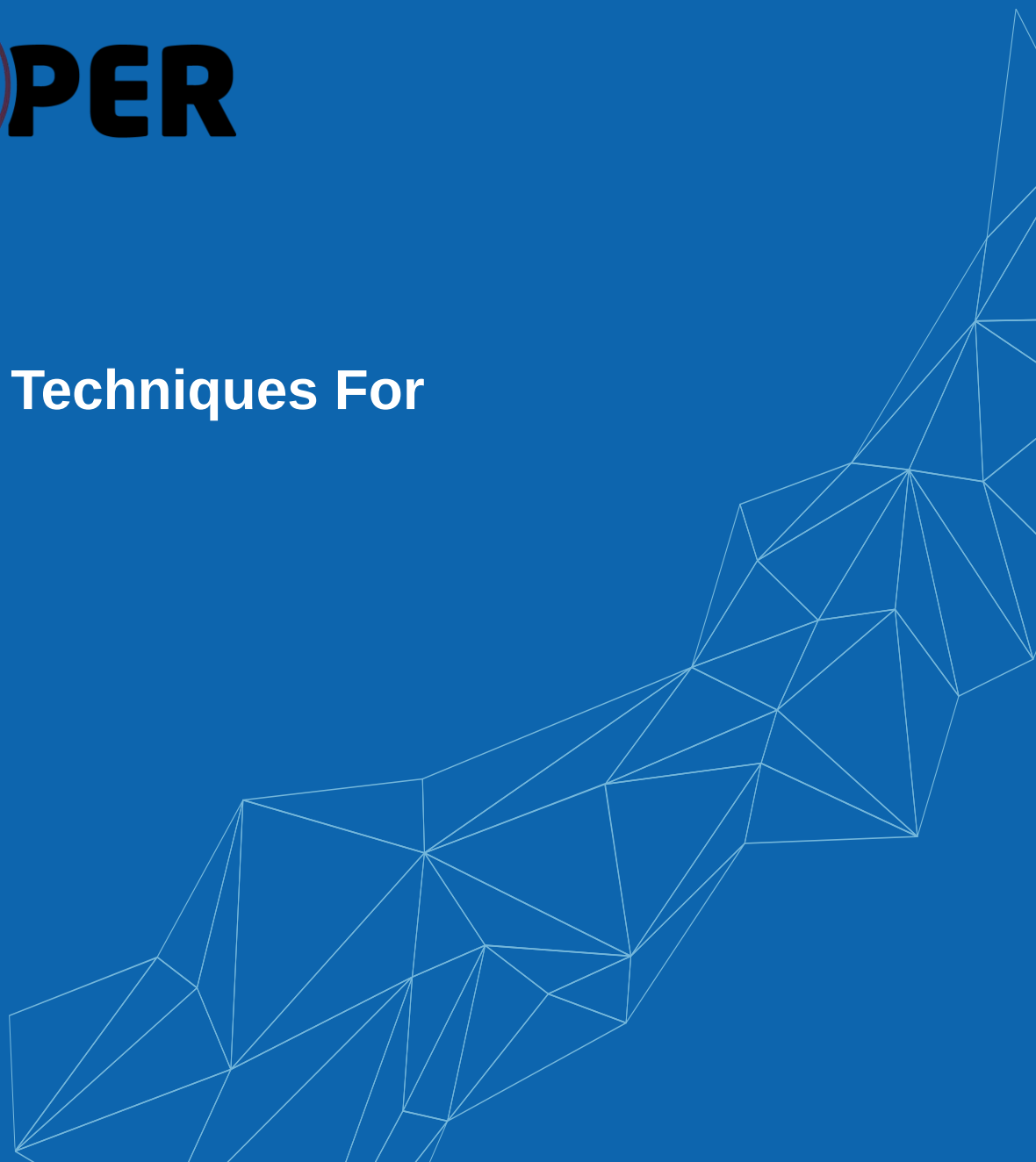
GREAT INDIAN DEVELOPER SUMMIT

Applying Regression Techniques For Predictive Analytics

Paviya George Chemparathy

April 2016



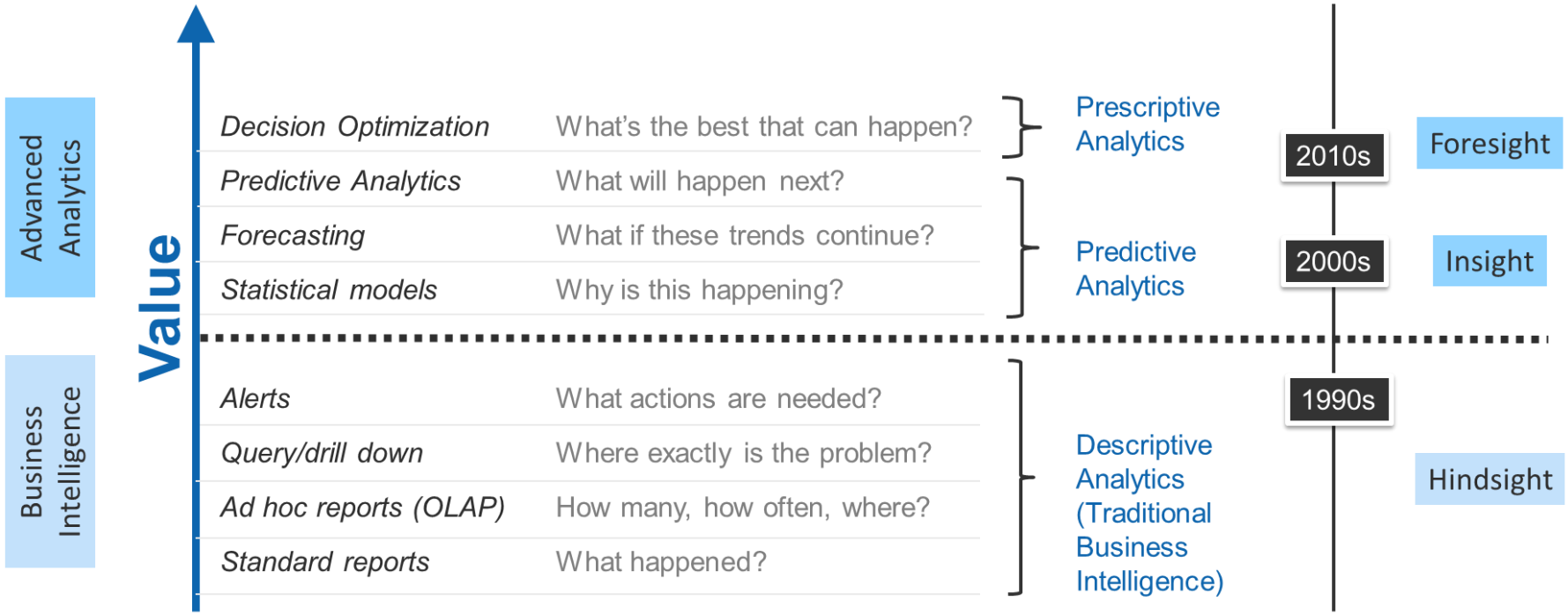
Sapient® Global Markets

AGENDA

1. Introduction
2. Use Cases
3. Popular Algorithms
4. Typical Approach
5. Case Study

Introduction

"Predictive Analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns."



Source: Competing on Analytics, The new Science of Winning

Predictive Analytics Use Cases

Popular use cases

Use Case Types	Examples	Analytics Technique
Predict a category	• What is the likely reason for a customer support call? • Which transactions are likely to be fraud?	Classification
Predict a value	• Energy demand forecasting • Predict next day stock price	Regression
Identify similar groups	• Customer Segmentation • News Clustering	Clustering
Co-occurrence Grouping	• What items are commonly purchased together? (Cross-selling opportunities)	Association analysis

Popular Algorithms

Algorithms for Predictive Analytics

	Supervised	Unsupervised
Continuous	**Regression** • Linear Regression • Polynomial Regression • Ridge, LASSO • ARIMA • SVR • Regression Trees	**Clustering** • K-Means • DBScan
Categorical	**Classification** • KNN • Decision Tree • Logistic Regression • Naïve Bayes • SVM • RandomForest • XGBoost	**Association Analysis** • Apriori • FP-Growth

Typical Approach

Data Understanding

Analyze and document available data and its relationships

Data Preparation

Construct a dataset suitable for subsequent phases

Modelling & Evaluation

Build and evaluate model

01

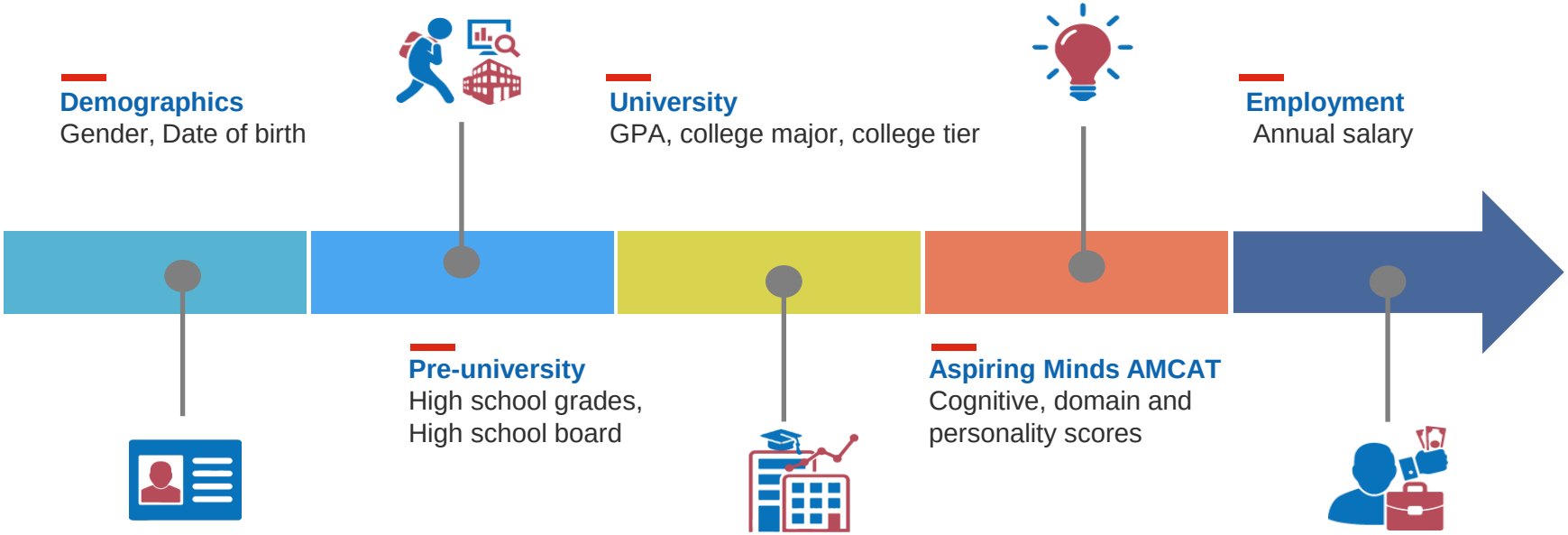
Case Study

- *Data Understanding*
- *Data Preparation*
- *Modelling*
- *Evaluation*

Case Study

India produces 1.5 million engineers every year. What determines the salary and the jobs these engineers are offered right after Graduation?

Can we predict salary from historic data?



Data Understanding

Examine summary characteristics

- Number of records, features, target variables

Identify problems

- Inaccurate or invalid values, missing values, unexpected distributions and outliers
- Using single variable summaries, categorical variable assessment & multiple variable summaries

Visualize data

- Histograms, Bar plot, Boxplots, Scatter plots
- R packages like ggplot & html widgets or tools like Rapidminer can be used

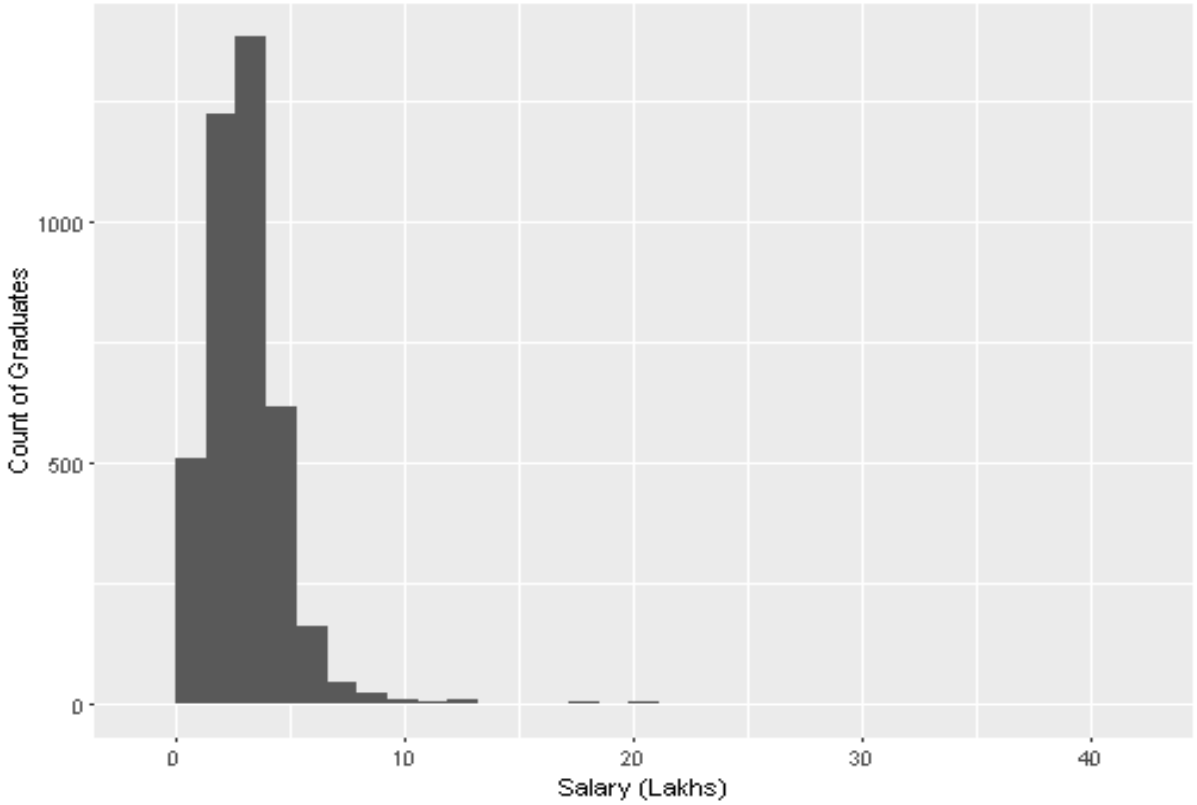
Why Data Understanding?

- Data Understanding helps us identify potential compatibility issues in the input data with predictive algorithms, so we can identify corrective measures for data preparation
- Each model makes several assumptions on the input data. For example, Multiple linear regression analysis makes the following assumptions:
  - Linear relationship between predictors and target variable
  - Predictors have a normal distribution
  - Predictors are not highly correlated
  - No cross-correlation with itself
  - Variance around the regression line is the same for all values of the predictors

Data Understanding

Example - Salary Statistics

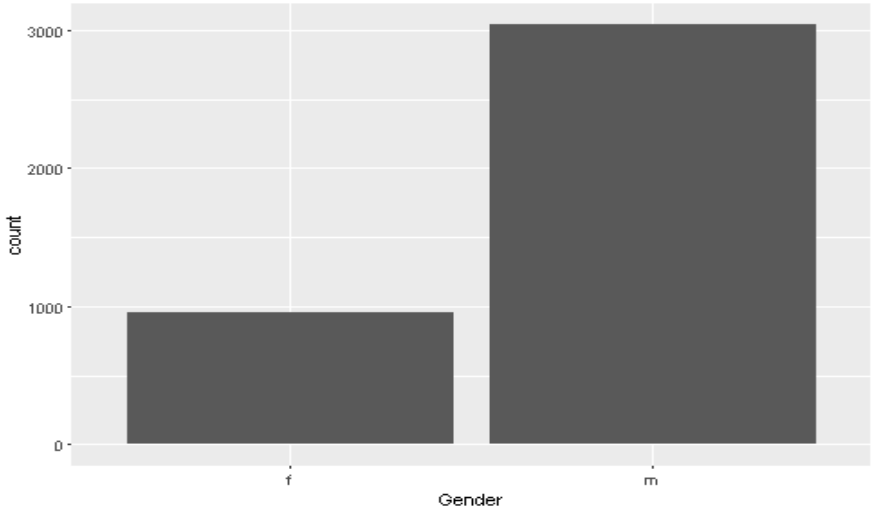
Salary is spread across a range of ₹35k - ₹40 lac and is right skewed.

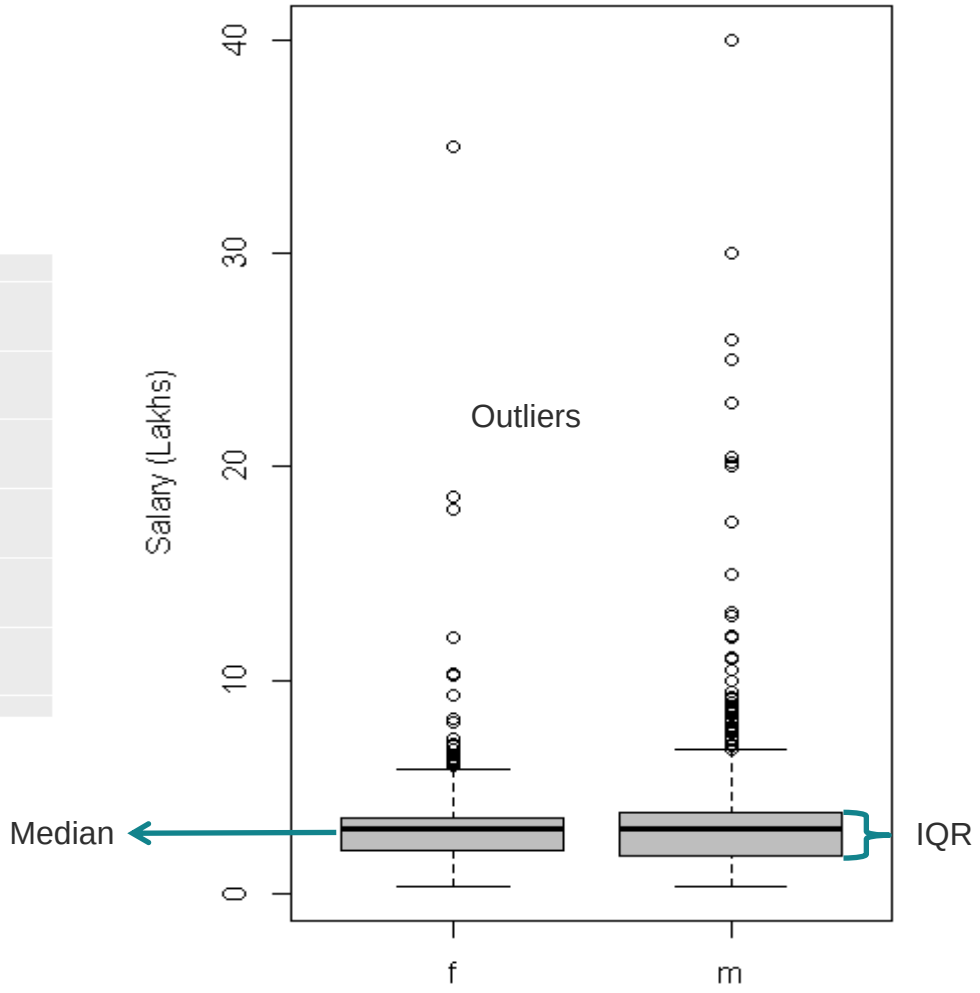


Min. Value	Ist Quartile	Median	Mean	3rd Quartile	Max Value
35,000	1,80,000	3,00,000	3,07,700	3,70,000	40,00,000

Data Understanding

Example - Salary by Gender





Data Preparation

Data preparation can be used to enrich, or standardize data.

The steps can be categorized into:

Collect Data

- Integrate data, de-normalize it into a dataset, collect it together

Pre-process

- Handle outliers
- Handle missing values (ignore, impute, drop)
- Normalize/scale
- Remove irrelevant and redundant variables

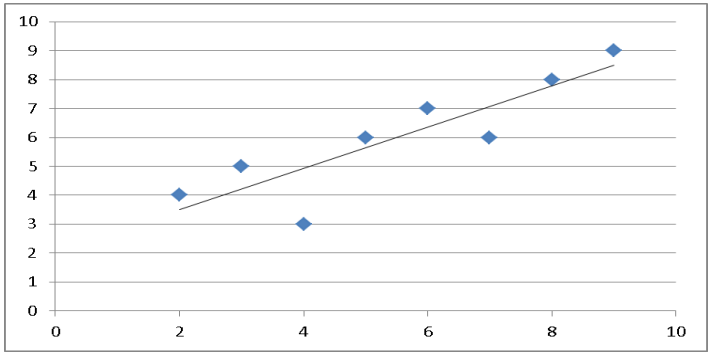
Transform/Enrich

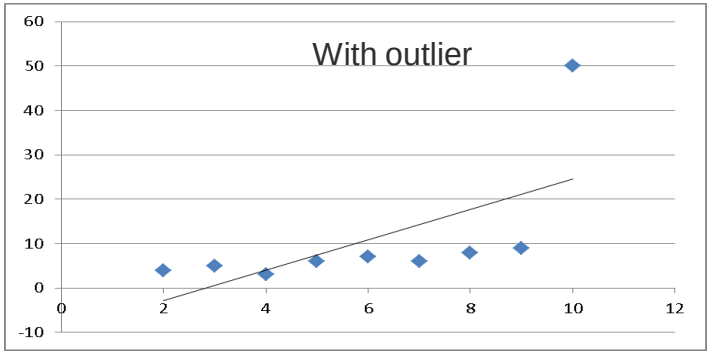
- Create derived/dummy variables
- Perform binning
- Correct skew

Why Data Preparation?

- To handle messy, inconsistent, or unstandardized data
- To ensure the data is presented to the algorithms in a way they can be used effectively

Let's take an example of outlier to understand the importance of data preparation





Data Preparation

Example –Binning , Dummy variables and Derived Variables

- **Perform Binning**
  Specialization column had more than 50 categories. We binned the categories into 6 main specializations – Computers, Electronics, Chemical, Mechanical, Civil, and Others.
  Similarly, 10th and 12th Board also had more than 200 categories which can be binned into 3 main boards – CBSE, ICSE and State Board.

10th CBSE	10th ICSE	10th State Board
0	0	1
1	0	0
1	0	0
0	1	0
1	0	0

- **Dummy variables**
  We created dummy variables for categorical features like specialization, degree, 10th Board, 12th Board and College state.

- **Derived variables**
  Examples of new features : Aggregate Percentile, Age, Improvement in 12th.

Data Preparation

Handle missing values

Missing data in the training data can reduce the power / fit of a model and can lead to a incorrect model. Missing data can be categorized into:

- **Missing completely at random**
- **Missing at random**
- **Missing that depends on unobserved predictors**
- **Missing that depends on the missing value itself**

Data Preparation

Example – Impute missing domain score

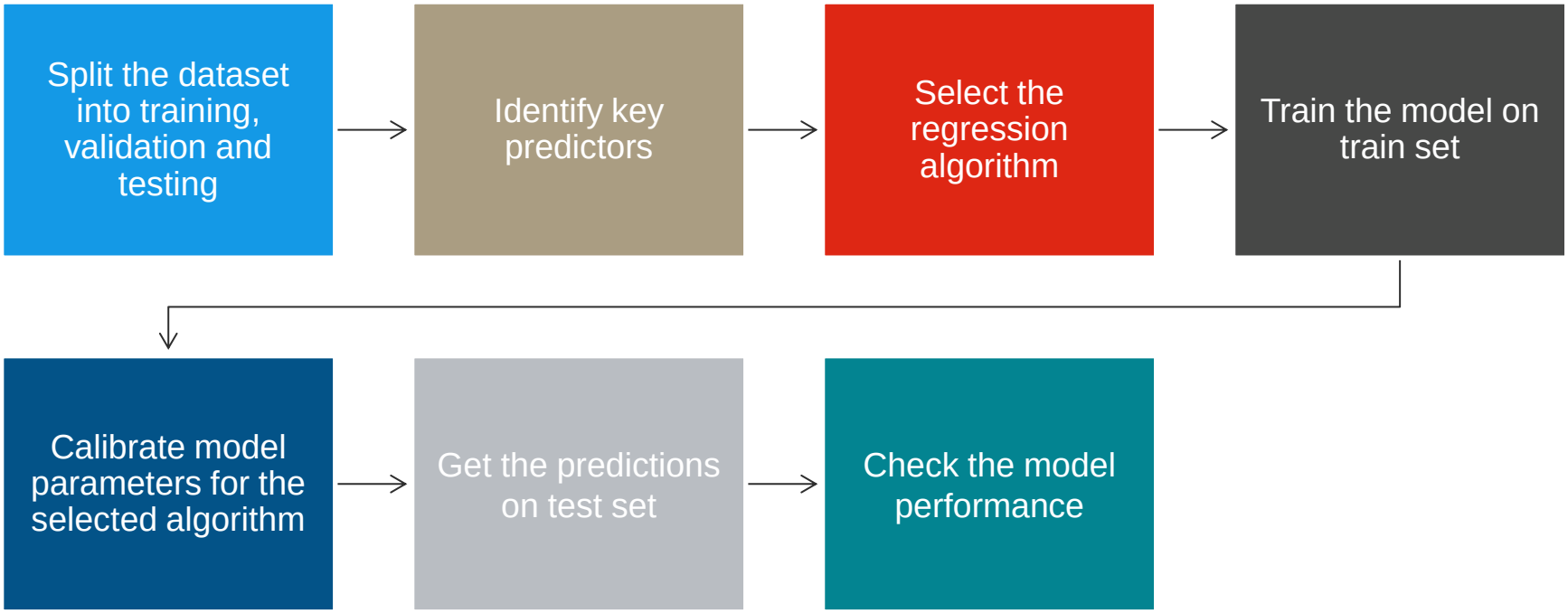
In the Salary data set the domain score had 256 missing values.

We applied the following steps to prepare the data:

- Analyzed the relation of domain with other features available
- Identified three features (Quant, logical and English) which had high correlation with the domain score
- Trained a linear regression model on the available data
- Predicted the missing domain values

Modeling Steps

The following diagram depicts key steps for building predictive models:



Modeling

Find predictors and train the models

For the salary dataset, we used the caret package and stepwise linear regression to identify the subset of good predictors.

We applied multiple regression algorithms like

- Linear Regression, SVR, Random Forest, MARS

Linear Regression Model

```
Coefficients:
          (Intercept)                 Quant         GraduationYear
           82430402.8                 143.8               -41012.8
           collegeGPA               English              X12percentage
               2290.7                 211.1                  1056.2
    conscientiousness            extraversion  openess_to_experience
             -10141.7                11585.2                 -7438.7

       X10percentage      agregatePecentile
              1205.9                 1423.9
        CollegeTier       CollegeCityTier
             -89898.2               -12942.3
```

MARS Model

```
                          coefficients
(Intercept)                  528892.6
CollegeTier                  -81000.3
h(74.5-X10percentage)          -632.9
h(X10percentage-74.5)          2479.1
h(62.9-collegeGPA)            -7484.3
h(collegeGPA-62.9)             1948.2
h(2010-GraduationYear)         -123.6
h(GraduationYear-2010)       -60151.6
h(GraduationYear-2013)        60309.5
h(265-English)                 1687.9
h(English-265)                  184.4
h(640-Logical)                  -80.1
h(Logical-640)                 1898.2
h(390-Quant)                    234.2
h(Quant-390)                    220.9
h(0.994051-Domain)           -55070.4
h(Domain-0.994051)        -14474562.2
```

Model

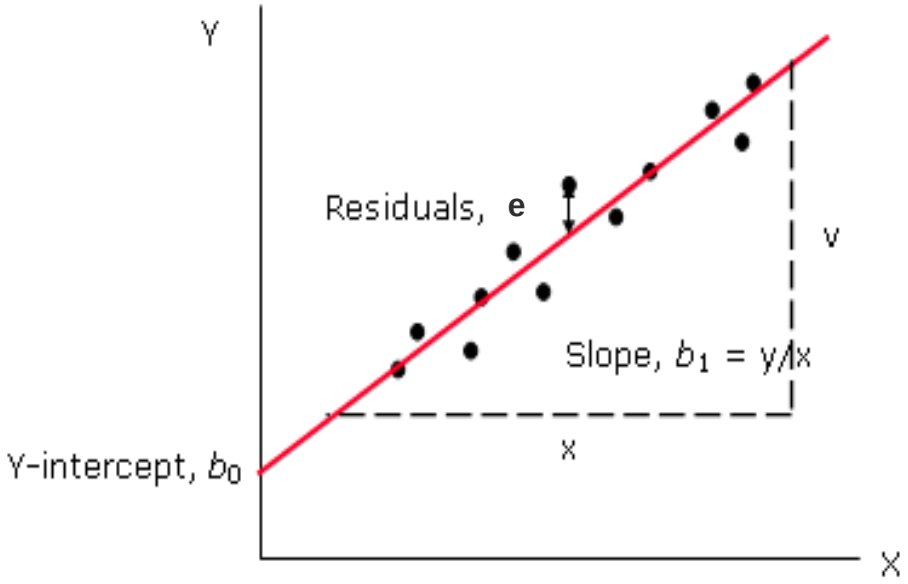
Linear Regression

In linear regression, we predict value of one variable from the value of a second variable. The variable we are predicting is called the target variable.

The variable we are basing our predictions on, is called the predictor variable.

The formula for a regression line is:

$Y' = b_0 + b_1X + e$



Let's take Salary and University GPA for the students. How we could predict a student's Salary if we knew his or her College GPA.

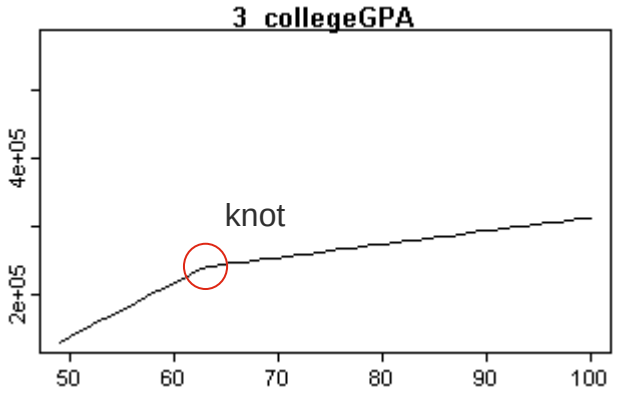
Salary = b_0 + b_1 * College GPA + e

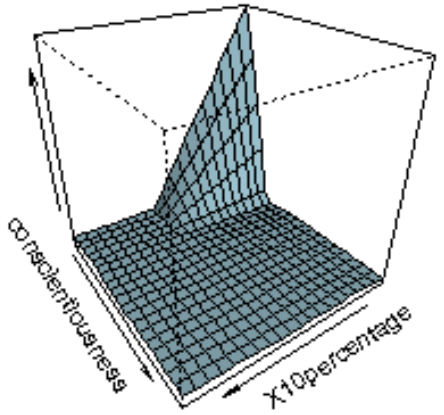
Model

Multivariate Adaptive Regression Splines (MARS)

MARS builds models of the form $\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x)$

The model is a weighted sum of basis functions multiplied by its coefficient.





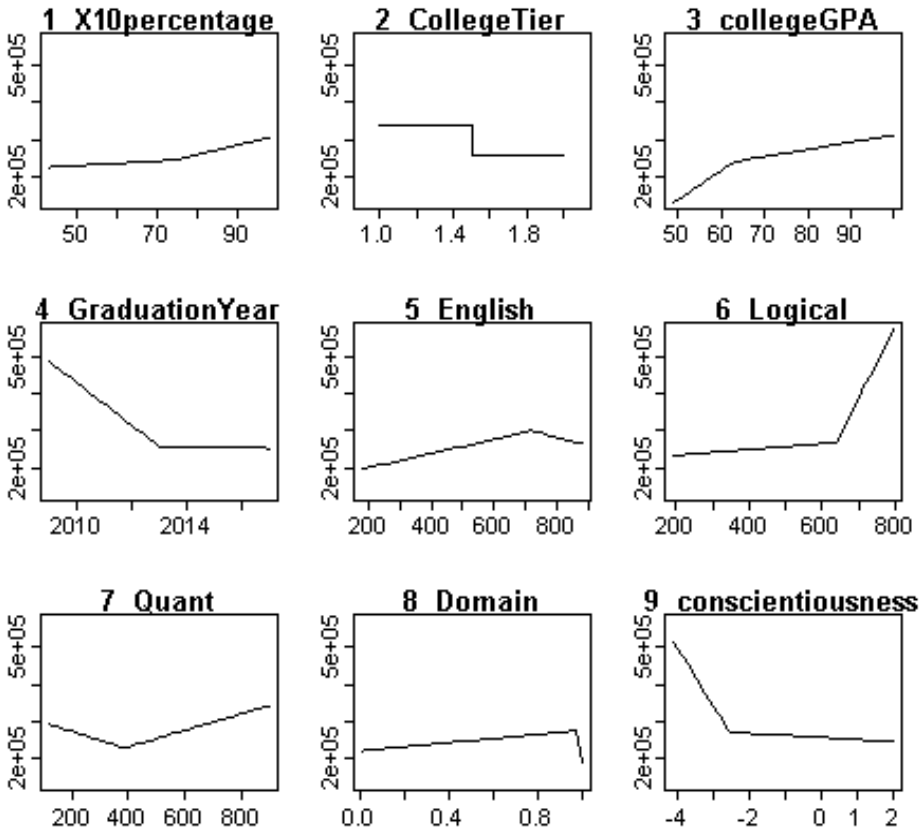
The MARS model for salary with College GPA as a predictor would look like

Salary = b_0 + b_1 * h(College GPA – 62.9) – b_2 * h(62.9 - College GPA) + e

Model

Features of MARS

- We used the mars implementation in R package Earth.
- MARS is a non-parametric regression.
- Break predictors into regions to permit nonlinearity.
- Handle Outliers better.
- Stepwise model building followed by a backwards elimination



Modeling

How to choose models

Choosing the correct model depends on the characteristics of your data. It's better to start with a simpler model. A few pointers while choosing the models are:

1. Is the relation between predictors and target variable Linear or non-linear?
2. Whether the data has outliers?
3. Is the data sparse?
4. Does the feature set has more numerical or categorical variables?
5. What is the distribution of the variables?
6. How complex the model can be?

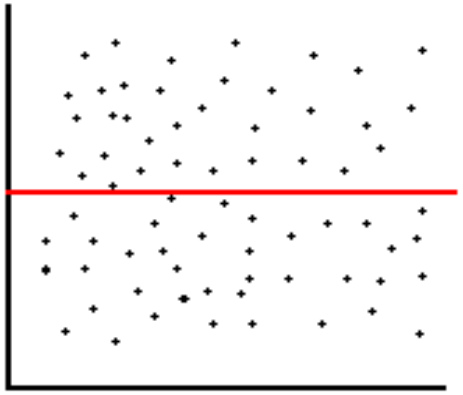
Model ensembling is a technique used to increase accuracy of ML as it reduces generalization error. It combines the predictions from multiple models to reduce over-fitting.

Evaluating the regression model

Each model has to be evaluated for the goodness of fit, unaccounted interactions, over-fitting, etc.

- Different evaluation parameters used for regression models include
  - Residual plots analysis, MSE, R-Squared

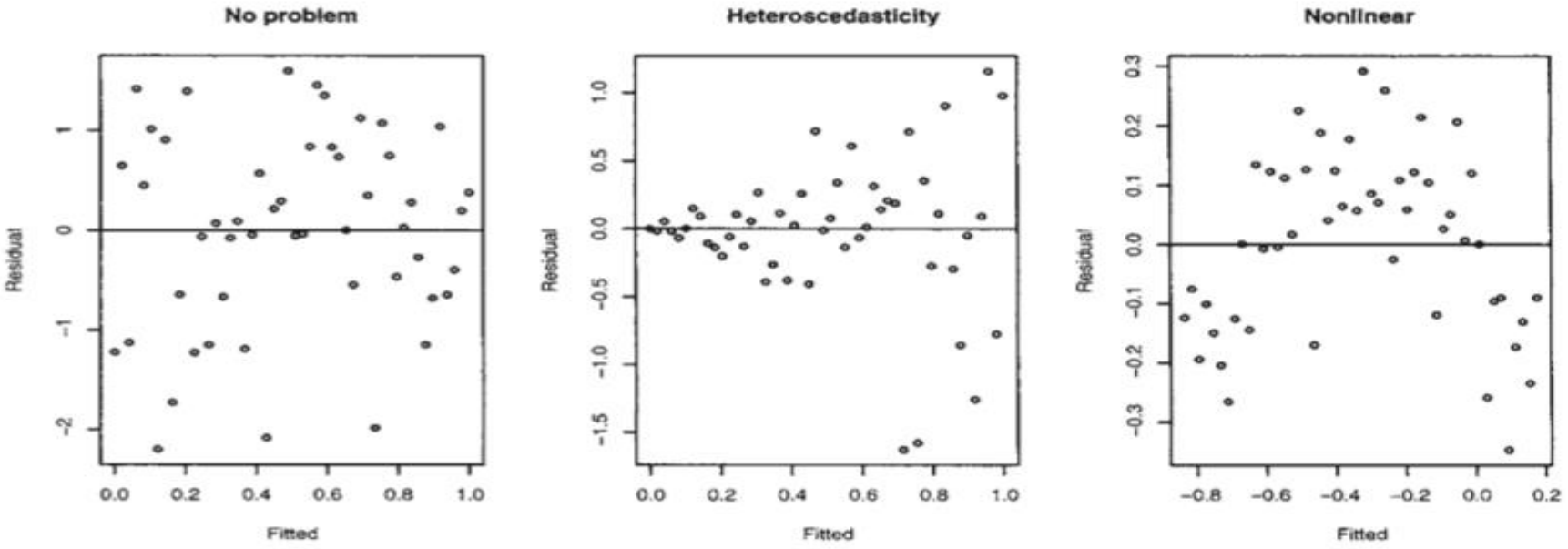
- Assumptions for the errors
  - the errors have mean zero
  - the errors are uncorrelated with each other
  - the errors are uncorrelated with each predictor



Ideal Residual plot

Evaluating the regression model

Residuals vs Fitted Plots



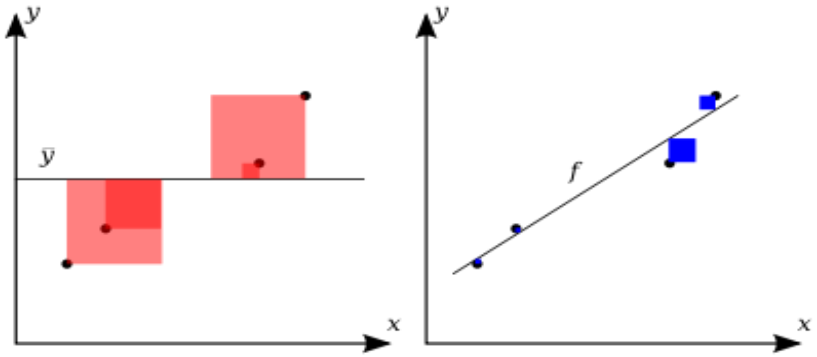
Evaluating the regression model

MSE and R-Squared

- Mean squared error (MSE) or Mean squared deviation (MSD) measures the average of the squares of the errors or deviations, that is, the difference between the estimator and what is estimated.

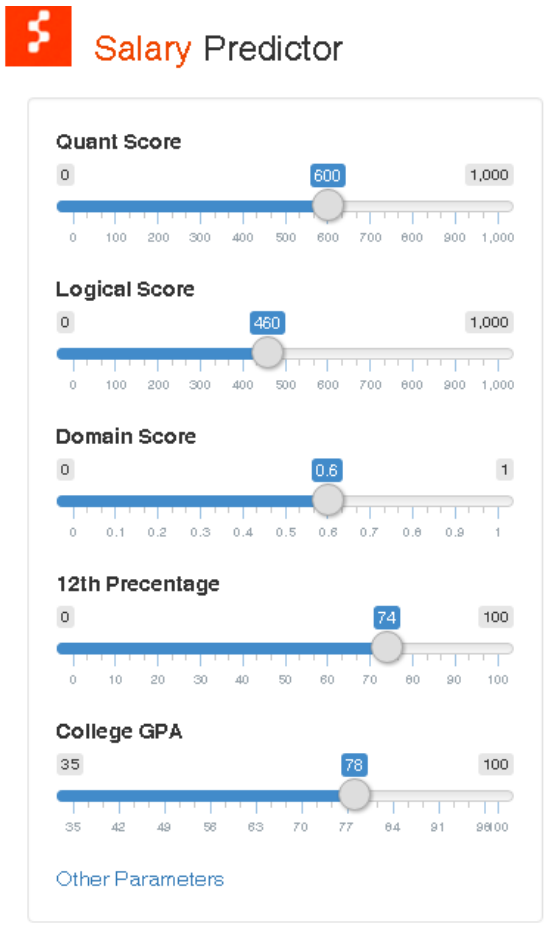
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

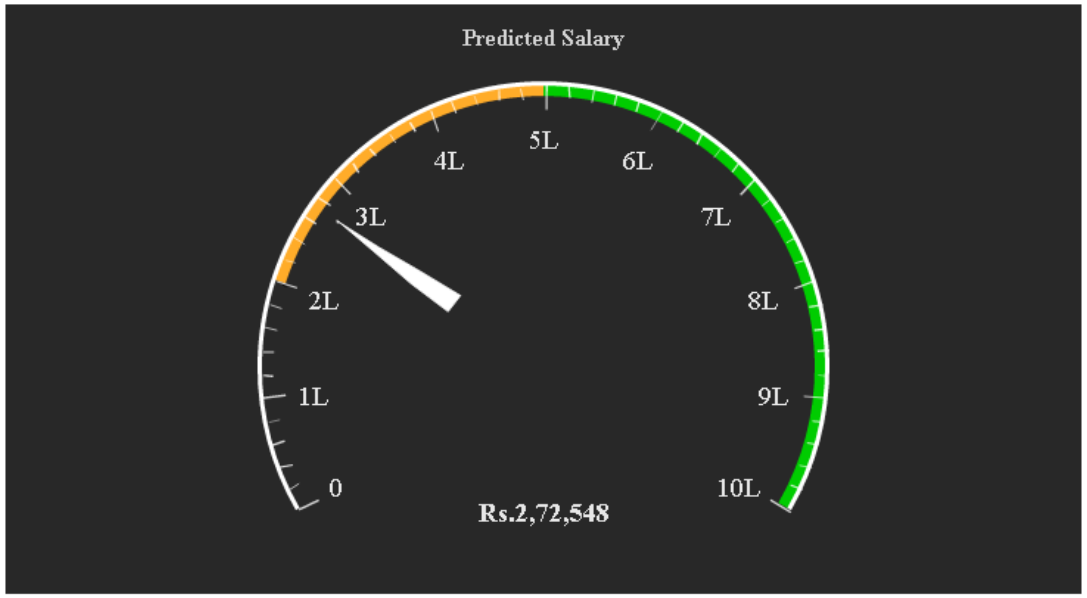
R squared indicates how well data fit a statistical model.
An R^2 of 1 indicates that the regression line perfectly fits the data, while an R^2 of 0 indicates that the line does not fit the data at all.



$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Salary Predictor





	Quant	Logical	Domain	X12percentage	collegeGPA	PredictedSalary
1	600	460	0.60	74	78	Rs.2,72,548

GREAT INDIAN DEVELOPER SUMMIT

THANK YOU

For more information please contact:
pchemparathy@sapient.com

Sapient® Global Markets