DA事前

# データ分析の概要

分析プロセスの代表例は下表の通りである

|  |  |
| --- | --- |
|  |  |
| 課題設定 | 解決したい問題を明確にし，データ分析の課題に落とし込む |
| データ収集 | 課題に必要なデータを集める |
| 探索的データ分析 | データの特徴や概要を掴み，分析の方針や仮説を立てる |
| モデル構築・評価 | モデルを作り，性能を評価しながら強化する |
| モデル運用 | 構築したモデルを実務に組み込んだり，分析から得られた知見を顧客に報告する |

これらは，戻りながら進めていく

# 課題設定・データ収集

データ分析の最初のプロセスは課題設定である．分析自体は目的ではないため，解決したい実務上の課題を明確にして，データ分析の課題に落とし込むことは重要なプロセスである．課題設定時のポイントは主に以下2点である．

* 回帰問題か分類問題に帰着できるか
  + 回帰：連続値（数値）を予測すること
  + 分類：離散値（クラス）を予測すること
* 必要なデータセットを用意できるか

《分析タスクの種類　kaggle本から》

《（できれば）課題設定の例》

# 探索的データ分析

## 探索的データ分析の概要

課題が設定できたら，分析の第一歩として，探索的データ分析（Explanatory Data Analysis，EDA）を行う．このプロセスの目的は諸説あるが，主なものとしては以下の通りである．

* データの傾向と特徴を理解する
* データ間の関係性を理解する
* データから不要な情報を取り除く

必ず順番通りに進むわけではなく，各プロセスを行き来しながら進める．

1. データの理解

データセットの背景やその分野の知識があるとは限らない．各データの意味が分からなければ，何が重要なデータが分からない．そこで，まずはデータセットの概要，データの意味，データの種類，値を確認する．そうすることで，後続プロセスでやるべきことが見えてくる．

1. データクリーニング

ほとんどの場合，データセットには欠損や分析に適さないデータが含まれている．そのまま分析してしまうと，解釈を誤るほか，モデル学習に影響を及ぼすことがある．また，分析しやすい形にすることも重要である．本章では，代表的な方法として，以下７つを紹介する．

|  |  |
| --- | --- |
| 方法 | 概要 |
| 欠損値の対応 | 欠損値を見つけて，削除等を行う |
| 外れ値の対応 | 外れ値を判定し，削除・クリッピング等を行う |
| 表記揺れの統一 | “男性”，”男”のような表記揺れを統一する |
| 重複データの削除 | 重複行，重複列を削除する |
| データの変換 | データの型変換，データを意味のある形へ変換 |
| 分類不可データの削除 | 分類できない，意味が無いデータの削除 |
| 正規化 | 列ごとに異なる数値のスケールを揃える |

1. 分析の切り口の検討

手法

* 基礎統計量の算出
* スライシング
* ドリルダウン
* 合成
* 尺度変換
* 正規化

1. 分析

データのクリーニングを行い，分析の切り口を決めたら，データの特徴や傾向，データ間の関係を理解するための分析を行う．分析を行う中で，疑問や仮説が生まれたら，再び前のプロセスを行い，分析の視点や切り口を検討する．本章では分析方法として以下6つを紹介する．

|  |  |
| --- | --- |
| 方法 | 概要 |
| 棒グラフ | データの差や大小を把握する |
| ヒストグラム | データの広がり，分布から特徴や傾向を掴む |
| 箱ひげ図 | データの位置や分布，外れ値を確認する |
| 折れ線グラフ | データの時系列での値の推移を確認する |
| 散布図 | ２変数間の相関，分布，外れ値を確認する |
| ヒートマップ | 多変数間の相関を確認する |

1. データの選択  
   データセットの特徴や傾向，目的変数との関連が把握できたら結果をまとめる．モデル学習時において必要なデータ・不要なデータに分け，必要なデータは目的変数との関連度合いに応じて優先度を付ける．モデル構築時には，優先度の高いデータに対して，重点的に特徴量生成や分析を行う．

## データの理解

確認する事項は以下の通り．

* データセットの概要
* データの意味
* データの種類
* 値の範囲
* それぞれの値の意味（特にカテゴリデータの場合）

確認したら，以下のような事柄について，仮説を考えてみる．

* 目的変数と関係がなさそうなデータ
* 優先的に分析をしたほうがよさそうなデータ
* まずは意味から調べる必要があるデータ
* データ間の関連性

あまりにも列数が多いデータに対しては，後述のヒートマップを使って，目的変数との相関が明らかに低いデータを分析対象から取り除くということもある．ただし，相関を見ているだけであるため，本当は関連がある可能性もあることに注意されたい．

データの種類については，基本の2パターンとして，質的データ/量的データに分類される．さらに，尺度分類により４つに分類される．

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 尺度分類 | 概要 | 可能な計算 | 主な統計量 |
| 質的データ （カテゴリデータ） | 名義尺度 | 他と区別するためのもの  （例：性別，血液型） | なし | 最頻値 |
| 順序尺度 | 順序には意味があるが，間隔には意味がない  （例：大学のGPA） | 大小比較のみ | 四分位数 |
| 量的データ （数値データ） | 間隔尺度 | 順序，差に意味はあるが，比率に意味はない．0は相対的なもの  （例：入試の偏差値） | 加算，減算 | 平均値  中央値  標準偏差 |
| 比例尺度 | 順序，差，比率に意味があり，0が絶対的な意味をもつ  （例：体重） | 四則演算 | 平均値  中央値  標準偏差  変動係数 |

分析においては，以下の点に注意する．

* 高水準から低水準のデータへの変換は可能だが，逆は不可能である
* データの種類に応じて可能な計算，算出できる統計量が変わってくる

## データクリーニング

|  |  |
| --- | --- |
| 方法 | 概要 |
| 欠損値の対応 | 欠損値を見つけて，削除等を行う |
| 外れ値の対応 | 外れ値を判定し，削除・クリッピング等を行う |
| 表記揺れの統一 | “男性”，”男”のような表記揺れを統一する |
| 重複データの削除 | 重複行，重複列を削除する |
| データの変換 | データの型変換，データを意味のある形へ変換 |
| 意味のないデータの削除 | 分類できない，意味が無いデータの削除 |
| 正規化 | 列ごとに異なる数値のスケールを揃える |

### 欠損値の対応

処理方法はさまざまであり，主な手法は以下の通り．

* 行ごと削除
* （数値データのみ）定数による補完
* （カテゴリデータのみ）固定カテゴリによる補完
* （カテゴリデータのみ）最頻値による補完
* 予測値による補完

＜欠損値の処理と削除をするコード例＞

### 外れ値の対応

外れ値の基準については，データの特性や分析目的に依存するため，可視化等を行った上で恣意的に除去することが多い．

＜可視化して外れ値を見つける図＞

しかしながら，１つずつ判断することが困難で，外れ値の閾値を定める必要があるときもある．閾値の決め方には様々なものがあるが，ここでは代表的なものとして，四分位範囲によるものを紹介する．

＜四分位範囲によるものの説明＞

### 表記揺れの統一

《今回，特に必要なければ，よくあるパターンを紹介するだけ》

### 重複データの削除

《今回の課題のデータでの例》

### データの変換

* カテゴリから数値への変換

### 意味のないデータの削除

＜参考書籍の例ではチケット番号．今回の課題データでの例を入れる＞

### 正規化

正規化とは，列ごとに異なる数値のスケールを揃えること．スケールが異なるとモデル構築で問題が発生する場合がある．最も一般的な手法は以下２点．

* 標準化：平均0，分散1に変換する
* Min-Maxスケーリング：最小値を0，最大値を1に変換する

使い分けについては，元データの特性や分布を考慮して使い分けられるのが理想だが，まずは，とびぬけた値が存在すれば標準化，存在しなければmin-maxという使い分けから始めると良い．

## 分析の切り口の検討

* 基礎統計量
* スライシング

データを様々な軸，水準で区切って層別化することで，意味付けされたそうの特徴を理解することができる，

＜今回の課題データでの例を図で示す＞

※注意点：層を分けすぎない

スライシングにおいて，層を細かく分けすぎると，本来は意味のない層でも意味のあるように見える分析結果となることがある．または，その逆で，本来意味のある層が意味のない層に見えてしまうこともある．

* ドリルダウン

ドリルダウンは，データを様々な軸で詳細化していく方法である．

＜今回のデータでの例＞

* 合成

合成は複数のデータから新たなデータを作成することである．やり方としては，データ同士を四則演算で組み合わせることが多い．

＜今回のデータでの例＞

* 尺度変換

データを低位に尺度変換する

* 正規化
* 無名数化
* 基礎統計量

## 分析

#### 単変数の分析

* ヒストグラム

#### 変数間の比較

* 棒グラフ（データ間の大小関係）
* 箱ひげ図（データの位置）
* バイオリン図（データ間の分布）

#### 相関分析

* 散布図
* ヒートマップ

# モデル構築・評価

## ポイントと概要

まずは，モデル構築における重要な2つの観点を述べる．

《ここの文言を考える》

* Point1：汎化性能

モデルを作成する主な目的は，未知のデータに対して予測を行ったり，そこから得られるモデル構造等から新たな知見を得ること．未知のデータに対して予測する性能を評価することが重要．

* Point2：解釈性

モデル構築・評価は以下の手順で行う．

《各項目にざっくりした説明を書く》

1. 評価指標の選択：
2. バリデーションの構築：汎化性能を評価し，モデル改良を行うための仕組みを作る．
3. 特徴量の生成，選択
4. モデル種類の選択
5. ハイパーパラメータチューニング
6. 学習・評価

## バリデーションの構築

《バリデーションとは，～すること．汎化性能を高めるために重要なプロセスである．》

バリデーションの方法は，大きく分けると以下２つである．

* ホールドアウト
* クロスバリデーション（交差検証）

## 評価指標の選択

当日の課題では分類問題を扱うため，分類モデルの評価指標を述べる．

《ここの情報収集が必要》

## 特徴量の生成と選択

《生成は書かない，選択方法を説明する》

## モデル種類の選択

《ここでは概要のみ述べて．詳細は付録》

### ロジスティック回帰

### 決定木

### ランダム・フォレスト

## ハイパーパラメータチューニング

* グリッドサーチ
* ランダムサーチ
* ベイズサーチ

## 評価

精度

解釈性

# モデル運用

顧客報告のフレームワークなど