データサイエンスの世界
～電気代はどうして高いの？～

株式会社テクノプロ　テクノプロ・デザイン社
名古屋支店
太田 征希

Confidential

2023/09/02

目次

TECHNOPRO Design

自己紹介

氏名：太田 征希

出身地：大阪府枚方市

趣味：バドミントン、コーヒー、ギター、
　　お酒を飲むこと、焚き火を見ること

経歴：

2022/03 北海道の某大学院(農学系)を修了

2022/04 テクノプロ・デザイン社に入社

2023/01 岐阜県の某メーカーに配属

データサイエンス未経験で入社
8ヶ月の研修を修了後、配属

現在のお仕事内容：

- お客さま先のデータを集計するツールの開発
- お客さま先社員向けのプログラミング関連教育の実施

1. 背景と目的

データサイエンスとは

統計学やプログラミングを活用してデータを解析し、有益な洞察を導き出す学問

データサイエンスが利用されている身近な例

1. 某大手回転寿司チェーン

   寿司皿にICチップを取り付け、「どの顧客がいつ何を何皿食べたのか」
   というデータを収集し、季節や天候によって変動する売れ筋ネタを特定した

2. 某食品メーカー

   ベビーフードに使われるダイスポテトの生産ラインに異常検知システムを導入し、
   不良品となるダイスポテトの自動選別を行なっている

データが溢れている現代において、データサイエンティストの需要は高まっている

本ブースで取り組むこと

データ分析を活用して電気代に影響を与える要因について考える

電気代は、住宅に関するさまざまな要因の影響を受けて変動する

電気代の分析を通して、データサイエンスがどんなものなのかを体験する

データの紹介

インドの住宅の電気代に関するオープンデータ

家電の有無、住人の数、集合住宅かどうか、電気代などのデータが含まれる

is_ac	is_tv	is_flat	ave_monthly_income	num_children	is_urban	amount_paid
あり	あり	集合住宅	9675.93	2	都市部でない	560.4814469
なし	あり	戸建	35064.79	1	都市部	633.2836786
あり	あり	集合住宅	22292.44	0	都市部でない	511.8791568
あり	あり	戸建	12139.08	0	都市部でない	332.9920353
なし	あり	戸建	17230.1	2	都市部	658.285625
なし	あり	集合住宅	24661.81	2	都市部	793.2423456
なし	あり	戸建	28184.43	1	都市部	570.3828451
なし	なし	集合住宅	16912.69	2	都市部	585.4051997
あり	なし		058.28	0	都市部	5
なし	なし		2545.5	2	都市部	8
なし	あり	戸建	15670.76	0	都市部でない	222.7345416
あり	あり	戸建	22527.33	0	都市部	606.1839763

説明変数 (feature)

目的変数 (target)

説明変数と目的変数の関係について分析する

2. 分析の流れ

全体の流れ

① 基礎集計

② 機械学習

③ 考察

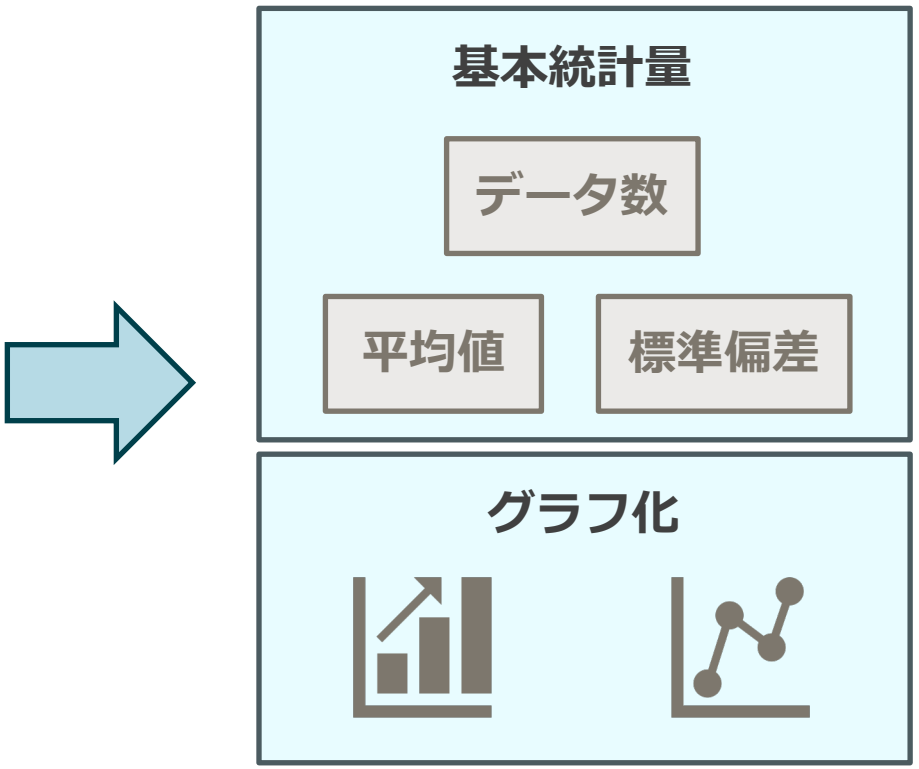
基礎集計と機械学習を用いて得たデータの知見について考察する

基礎集計とは

データの傾向を事前に把握する作業

テーブルデータ

	urban	amount_paid
2	都市部でない	560.4814469
1	都市部	633.2836786
0	都市部でない	511.8791568
0	都市部でない	332.9920353
2	都市部	658.285625
2	都市部	793.2423456
1	都市部	570.3828451
2	都市部	585.4051997
0	都市部	653.2008685
2	都市部	606.015138
0	都市部でない	222.7345416
0	都市部	606.1839763

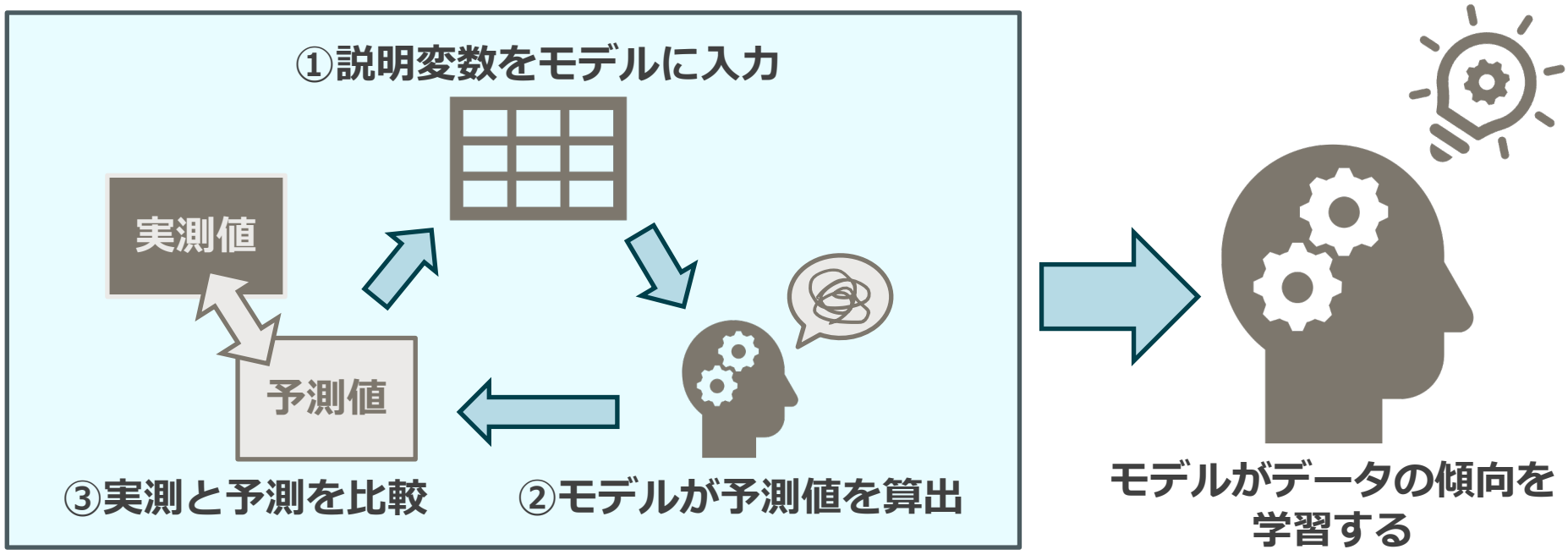


データをグラフ化したり基本統計量を算出し、データの特徴を確認する

機械学習とは

入力された説明変数に対する目的変数の予測(=予測値)を出力する仕組み
データを用いて作成した「機械学習モデル」を用いて予測値を算出する

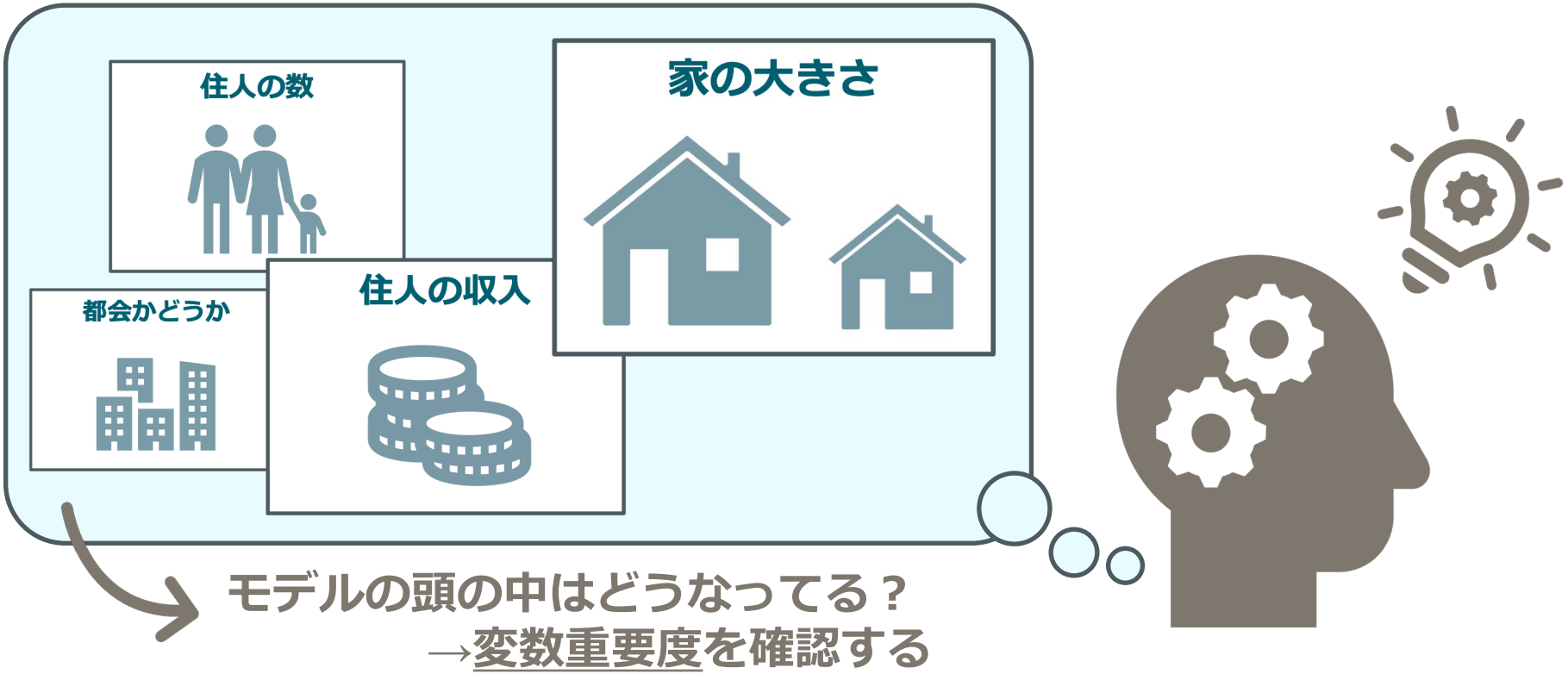
機械学習モデル構築の”イメージ” (実際のアルゴリズムとは異なります)



データの特徴を反映した機械学習モデルを作成し、推論に活用する

今回の分析で機械学習をどう活用するのか

モデルが予測値を算出する上で重要視した説明変数を確認する



モデルの変数重要度を参考に、電気代に影響を与える要因について考える

アイスブレイク

どの説明変数が電気代に最も大きな影響を与える？

実際に考えてみましょう（仮説立案）

5min

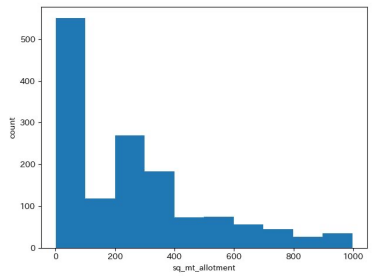
ハンズオン

3. 基礎集計

データの特徴の確認

アイスブレイクで上げられたデータ項目について基本統計量を確認し、グラフ化する

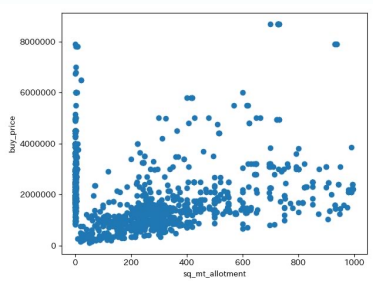
数値データのグラフ化



ヒストグラム

横軸: データの階級
縦軸: データの数

階級ごとの頻度分布を確認する

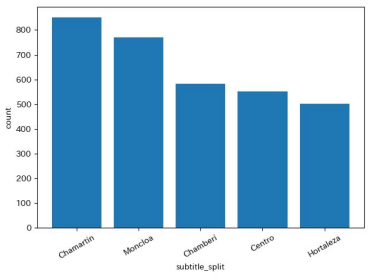


散布図

横軸: 確認する説明変数
縦軸: 電気代

説明変数と電気代の相関を確認する

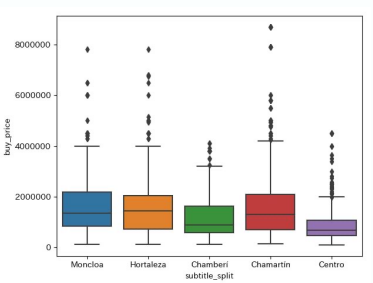
カテゴリデータのグラフ化



棒グラフ

横軸: データのカテゴリ
縦軸: データの数

カテゴリごとの頻度分布を確認する



箱ひげ図

横軸: データのカテゴリ
縦軸: 電気代

カテゴリごとの電気代の分布を確認する

データの特徴を確認し、アイスブレイクで立てた仮説が正しいと言えそうか確かめる

4. 機械学習

機械学習の流れ

① モデルの学習

全体の8割のデータを用いてデータの特徴をモデルに学習させる

② モデルの評価

全体の2割のデータを用いてモデルの予測精度を確認する

③ 変数重要度の確認

変数重要度を可視化し、モデルがどの説明変数を重要視して
予測値を算出したのかを確かめる

モデルを構築・評価した上で、変数重要度を確認する

TECHNOPRO Design

モデルの学習

「LightGBM」と呼ばれる仕組みを用いて機械学習モデルを構築する

LightGBMの特徴

1. 学習にかかる時間が短く、かつ精度が高い
2. 変数重要度を確認できる

→機械学習モデルを構築する上でまず初めに使われる手法

LightGBMの使い方

Pythonのライブラリ（追加機能のようなもの）に含まれる関数を呼び出し、
関数にデータを入力して学習させる

LightGBMにデータを入力し、データの特徴を学習させる

TECHNOPRO Design

モデルの評価

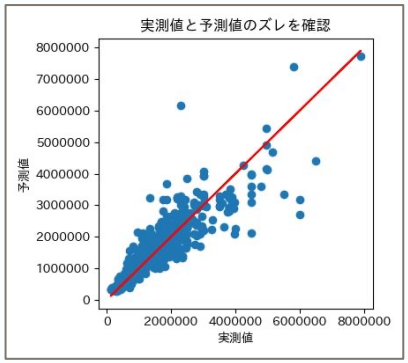
予測値と実測値を確認し、評価指標の値を確認する

【評価指標の確認】
RMSE : 487888.15473215905
R2 : 0.8002396269204338

	実測値	予測値
321	1075000	1354551
2143	2700000	2420420
1619	350000	652764
1565	2595000	2763556

1. 評価指標の確認と実測値・予測値の比較

定量的なモデルの評価指標(RMSEと決定係数)を算出し、
実測値とモデルの予測値を見比べて
どれくらいモデルの予測が正しいのかを確認する



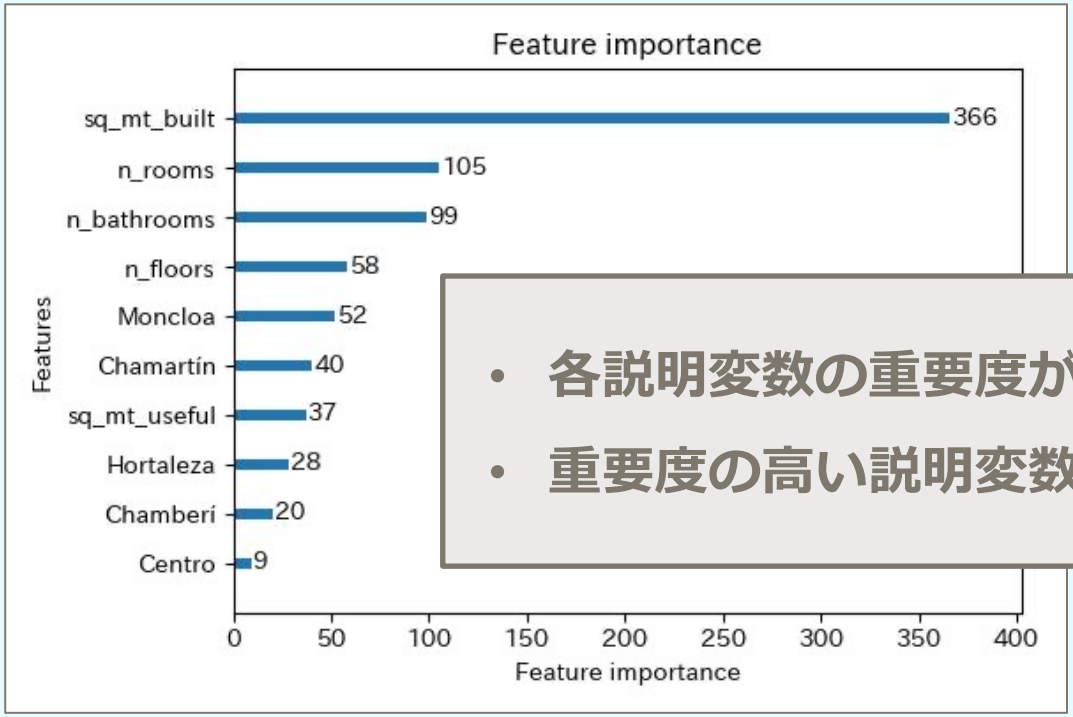
2. yyplotの確認

横軸に電気代の実測値、縦軸に予測値をとり、
実測値と予測値にどれくらい乖離があるのかを
視覚的に確認する

モデルの予測が信用できるのかを判断する

変数重要度の確認

変数重要度を表現した横棒グラフを描画する



- 各説明変数の重要度が棒の長さで表現される
- 重要度の高い説明変数が上から順に並べられる

作成したモデルが、予測値を算出する上でどの説明変数を重要視したのかを確認する

5. 考察

今回作成したモデルは「子供の人数」を重要視していた

なぜ「子供の人数」が電気代に影響を与える？

考えられる理由

- 子供は大人と比べて家にいる時間が長い
- 子供の面倒を見るために親も家にいる時間が長くなる
- 親は子供に気を使うので、空調の使用を惜しまない

⇒ 「家に人がいる時間の長さ」や「1日あたりのエアコンの稼働時間」と「子供の人数」の関係性について調べると、電気代が高くなる要因について理解が深まりそう

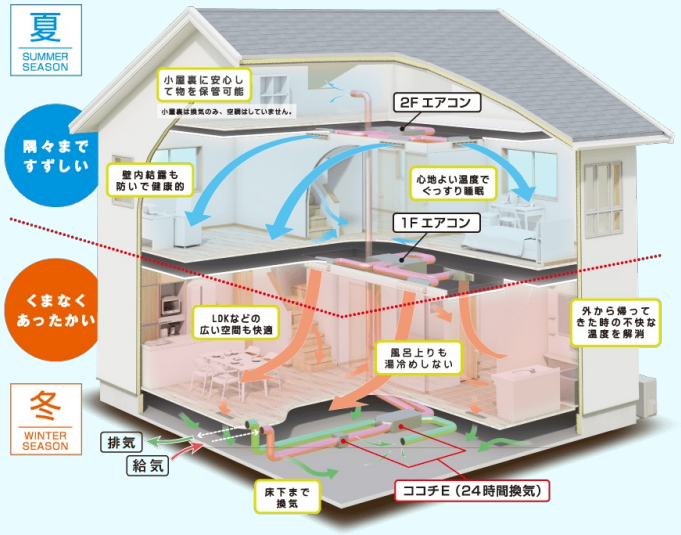
分析で得た知見をもとに再び仮説を立て、さらに分析を進めていく

おわりに

電気代を抑えられる住宅のカタチ

室内の空気を効率よく循環させる「Z空調」な住宅

Z空調な住宅とは？



- 全館空調で家のどこにいても快適
- 機密性・断熱性が高く、
  熱効率が良い
- 夏は上からの風を送り、
  冬は床から空気の流れを作ることで
  家全体の温度を効率よく調節する
- 24時間稼働させても電気代が安い

「今話題のZ空調は何がすごいのか？桧家住宅のZ空調の特徴・メリットを解説」https://www.hinokiya-woods.com/column/1012/

「Z空調」な住宅が、コスト効率良く快適な環境を提供します

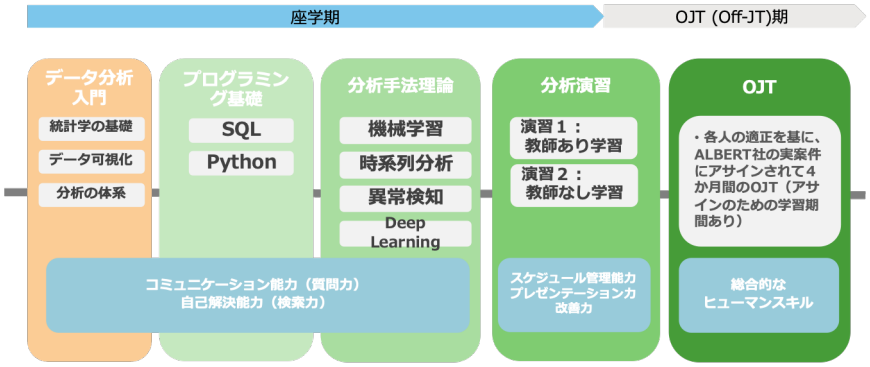
おわりに

テクノプロでデータサイエンティストになるには

戦略研修採用

- カリキュラムに沿って研修を受講し、データサイエンスの知識を身につける
- 戦略研修採用に合格することが必要

戦略研修のカリキュラム



社内公募

- パッケージ研修を受講し、データサイエンスの知識を身につける
- 入社後、自己実現制度に応募する

パッケージ研修のカリキュラム（一部）

No.	研修名
1	**ロジカル・シンキング基本コース** ロジックツリー、MECE、ピラミッドストラクチャーなど、ロジカル・シンキングの基本的な知識と、論理思考の実践的な使い方を、例題を解きながら身につけます。
2	**【Aidemy】統計学基礎** データ分析の基礎となる統計学を初学者の方でも学習を始められるように動画にしてわかりやすく説明します。 変数、グラフ、相関係数など。
3	**【Aidemy】統計学標準** 時系列データの取扱いから線形回帰モデルの分析までを取り扱っています。 この講座を見ることで実務でも役立つ統計学を学ぶことができます。
4	**【Aidemy】SQL基礎** データベースからの読み出し、データベースへの書き込み等の基礎的なSQL文法を、実際に記述し、実行しながら身につけていきます。
5	**【Aidemy】Python入門** 機械学習で最も使われるプログラミング言語「Python」の基礎を学びます。文字の出力、変数の概要、条件分岐、ループなど、「Python」の基本的な使い方をマスターしましょう。
6	**【オンライン】Pythonプログラミング**

実務経験の有無を問わず、データサイエンティストになれる環境が用意されています

Appendix

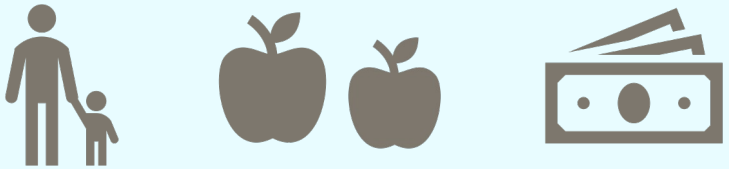
数値データとカテゴリデータ

データは数値データ（量的データ）とカテゴリデータ（質的データ）に分けられる

数値データ（量的データ）

- 数字で定量的に表すことができる
- 値の差に意味を持つ
  （数値の差が持つ意味が等しい）

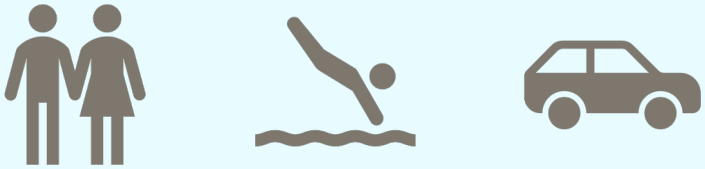
ex.
年齢、物の数、面積、収入、etc



カテゴリデータ（質的データ）

- 数字で定量的に表せない
- データを分類したり、
  種類を区別するためのデータ

ex.
性別、順位、車のナンバー、etc



データ型によって扱い方が異なる

データ項目一覧表

データ項目名	説明	データ型
num_rooms	住宅の部屋数	数値
num_people	世帯の住人数	数値
housearea	家の面積	数値
is_ac	エアコンの有無	カテゴリ
is_tv	テレビの有無	カテゴリ
is_flat	集合住宅かどうか	カテゴリ
ave_monthly_income	世帯の平均月収	数値
num_children	世帯の子供の人数	数値
is_urban	市街地にある住宅かどうか	カテゴリ
amount_paid	毎月の電気代（目的変数）	数値

予測精度の評価指標

本分析では評価指標としてRMSEと決定係数を用いる

RMSE

- 実測値と予測値にどれくらいズレがあったのかを示す指標
- 予測を大きく外した時に大きなペナルティを与える
- 小さいほど良い

$$RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

決定係数（R2）

- データに対するモデルの当てはまりの良さを示す
- モデルがデータの特徴をどれくらい捉えているかを示す
- 0から1の値をとり、1に近づくほど当てはまりが良い

$$R^2 = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y}_i)^2}$$

y_i: 実際の値, $\hat{y}_i$: 予測値, $\bar{y}_i$: 平均値, n: データの総数