統計コンペ (Kaggle) 報告

TesnorFlow と Keras による NN の実装

中川 浩輔

応用数理学科

January 9, 2025

Contents

Contents

Kaggle とは

世界中のデータサイエンティストや機械学習エンジニアが集まる、データ分析コンペティションおよび機械学習モデルの共有プラットフォーム。企業や研究機関が抱える実世界の課題が、データセットとともに公開され、参加者は自身のスキルを駆使して最適なモデルを構築し、競い合う。

- データセット: 多種多様な分野の膨大なデータセットが公開され、自由に利用可。
- コミュニティ: フォーラムやディスカッションを通じて、他のユーザーと知識を共有し、スキルアップが可能。
- 学習教材: チュートリアルやコースが豊富で、初心者から上級者まで学ぶことができる。
- チーム参加も可能。
- さまざまな賞金や景品

参加した Competiton

Regression with an Insurance Dataset (回帰による保険料予測)

テーブルデータに含まれるさまざまな特徴量から、保険料を予測するための回帰モデルを開発する。

120 万人分の学習データと 80 万人分のテストデータが与えられる。カテゴリ、数値、テキストが混在し、実際に直面する複雑さを模倣するために、欠損値、不正なデータタイプ、偏った分布が含まれる。
予測のターゲット変数は「Premium Amount」。

評価指標は $RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(\hat{y}+1) - \log(y+1))^2}$

データセットの特徴

Age: 被保険者の年齢
Gender: 被保険者の性別 (カテゴリ: Male, Female)
Annual Income: 被保険者の年収（歪曲）
Marital Status: 被保険者の婚姻状況 (カテゴリ: Single, Married , Divorced)
Number of Dependents: 扶養家族の数 (欠損値あり)
Education Level: 取得した最高教育レベル (カテゴリ: High School's, Master's, PhD)
Occupation: 被保険者の職業（カテゴリ: Employed, Self-Employed, Unemployed）
Health Score: 健康状態を表すスコア
Location: 場所の種類 (カテゴリ: Urban, Suburban, Rural)
Policy Type: 保険種別 (カテゴリ: Basic, Comprehensive, Premium)

Previous Claims: 過去の請求件数 (外れ値あり)
Vehicle Age: 保険対象車両の年齢
Credit Score: 被保険者の信用スコア (欠損値あり)
Insurance Duration: 保険契約の期間 (年数)
Premium Amount: 保険料額を表すターゲット目的変数 (歪曲)
Policy Start Date: 保険契約の開始日 (YYYY-MM-DD 形式)
Customer Feedback: 顧客からの短いフィードバックコメント (カテゴリ: Good, Average, Poor)
Smoking Status: 被保険者の喫煙状況 (カテゴリ: Yes, No)
Exercise Frequency: 運動の頻度 (カテゴリ: Daily, Weekly, Monthly, Rarely)
Property Type: 所有する不動産の種類 (カテゴリ: Condo, House, Apartment)

id	Age	Gender	Annual Income	Marital Status	Number of Dependents	Education Level	Occupation	Health Score	Location	Policy Type	Previous Claims	Vehicle Age	Credit Score	Insurance Duration	Policy Start Date	Customer Feedback	Smoking Status	Exercise Frequency	Property Type	Premium Amount
0	19.0	Female	10049.0	Married	1.0	Bachelor's	Self-Employed	22.59876067181390	Urban	Premium	2.0	17.0	372.0	5.0	2023-12-23 15:21:39.134960	Poor	No	Weekly	House	2869.0
1	39.0	Female	31678.0	Divorced	3.0	Master's		15.569730989408000	Rural	Comprehensive	1.0	12.0	694.0	2.0	2023-06-12 15:21:39.111551	Average	Yes	Monthly	House	1483.0
2	23.0	Male	25602.0	Divorced	3.0	High School	Self-Employed	47.17754928786460	Suburban	Premium	1.0	14.0		3.0	2023-09-30 15:21:39.221386	Good	Yes	Weekly	House	567.0
3	21.0	Male	141855.0	Married	2.0	Bachelor's		10.938144158664600	Rural	Basic	1.0	0.0	367.0	1.0	2024-06-12 15:21:39.226954	Poor	Yes	Daily	Apartment	765.0
4	21.0	Male	39651.0	Single	1.0	Bachelor's	Self-Employed	20.376093627736900	Rural	Premium	0.0	8.0	598.0	4.0	2021-12-01 15:21:39.252145	Poor	Yes	Weekly	House	2022.0
5	29.0	Male	45963.0	Married	1.0	Bachelor's		33.05319768402280	Urban	Premium	2.0	4.0	614.0	5.0	2022-05-20 15:21:39.207847	Average	No	Weekly	House	3202.0
6	41.0	Male	40336.0	Married	0.0	PhD			Rural	Basic	2.0	8.0	807.0	6.0	2020-02-21 15:21:39.219432	Poor	No	Weekly	House	439.0
7	48.0	Female	127237.0	Divorced	2.0	High School	Employed	5.769783092512090	Suburban	Comprehensive	1.0	11.0	398.0	5.0	2022-08-08 15:21:39.181605	Average	No	Rarely	Condo	111.0
8	21.0	Male	1733.0	Divorced	3.0	Bachelor's		17.869550814826300	Urban	Premium	1.0	10.0	685.0	8.0	2020-12-14 15:21:39.198406	Average	No	Monthly	Condo	213.0
9	44.0	Male	52447.0	Married	2.0	Master's	Employed	20.47371785695600	Urban	Comprehensive	1.0	9.0	635.0	3.0	2020-08-02 15:21:39.144722	Poor	No	Daily	Condo	64.0
10	56.0	Female	8054.0	Married	1.0	Bachelor's	Unemployed		Rural	Premium	1.0	8.0	431.0	8.0	2021-05-12 15:21:39.134960	Average	No	Monthly	Condo	857.0
11	23.0	Male	30983.0	Single	3.0	Master's		5.813128940949040	Urban	Premium		6.0	597.0	8.0	2020-03-22 15:21:39.155231	Good	No	Rarely	Condo	1447.0
12	25.0	Female	23706.0	Single	4.0	Master's	Employed	4.090538023921370	Urban	Comprehensive	2.0	19.0		2.0	2021-10-28 15:21:39.086123	Poor	Yes	Weekly	Apartment	703.0
13	44.0	Female	70893.0	Single	0.0	Bachelor's		55.89632239159920	Suburban	Premium	0.0	3.0	511.0	6.0	2023-01-20 15:21:39.132191	Good	Yes	Weekly	House	1847.0
14	40.0	Female	23897.0	Divorced		High School	Self-Employed	29.08203615642360	Suburban	Basic	2.0	15.0	498.0	1.0	2021-08-14 15:21:39.233998	Good	No	Rarely	Condo	30.0
15	18.0	Male	6076.0	Married	2.0	High School	Employed	7.442964015746720	Urban	Premium	1.0	12.0	584.0	5.0	2020-10-17 15:21:39.209847	Good	Yes	Monthly	Apartment	849.0
16	59.0	Female	28266.0	Divorced	2.0	PhD	Unemployed	21.67346054211260	Urban	Basic	0.0	16.0		3.0	2024-04-22 15:21:39.255696	Average	Yes	Weekly	Condo	183.0
17	34.0	Female	45907.0	Divorced	4.0	High School	Self-Employed	24.0596387644636	Suburban	Comprehensive	0.0	0.0	694.0	8.0	2020-12-27 15:21:39.090124	Poor	Yes	Daily	Condo	643.0
18	18.0	Male	29071.0	Married		Bachelor's	Employed	20.8389765352726	Rural	Premium	1.0	5.0		3.0	2021-10-13 15:21:39.094752	Poor	Yes	Rarely	Apartment	787.0
19	40.0	Female	123751.0	Single	2.0	Master's	Self-Employed	24.95531647911230	Suburban	Premium	0.0	8.0	420.0	2.0	2020-06-25 15:21:39.152732	Good	Yes	Rarely	Condo	40.0
20	39.0	Male	6837.0	Divorced	2.0	High School	Self-Employed	17.814458917250200	Urban	Premium		7.0	595.0	2.0	2024-06-25 15:21:39.272456	Good	Yes	Daily	Apartment	1204.0
21	44.0	Female	14042.0	Divorced	2.0	Master's		4.551825118494740	Urban	Basic		6.0	799.0	5.0	2023-07-31 15:21:39.237118	Good	Yes	Weekly	Apartment	2670.0
22	22.0	Male		Divorced	4.0	PhD		25.58378995160540	Urban	Comprehensive		5.0	773.0	5.0	2021-10-09 15:21:39.258696	Good	Yes	Monthly	House	202.0
23	34.0	Female	32762.0	Divorced	1.0	Bachelor's		21.96760938053780	Rural	Premium	0.0	10.0	798.0	1.0	2020-01-16 15:21:39.096739	Average	Yes	Monthly	Condo	3869.0
24	46.0	Male	24708.0	Married	3.0	High School	Employed	26.3412591810326	Urban	Premium		0.0	543.0	4.0	2021-08-30 15:21:39.134960	Average	No	Daily	House	1136.0
25	19.0	Male	72482.0	Married	2.0	Bachelor's	Unemployed	27.27941569575120	Suburban	Premium	0.0	17.0	416.0	3.0	2021-03-10 15:21:39.097737	Good	Yes	Weekly	Condo	671.0
26	49.0	Female	82584.0	Married	0.0	Bachelor's	Unemployed	23.86738827838690	Rural	Comprehensive		10.0	425.0	9.0	2019-10-06 15:21:39.091080	Average	Yes	Weekly	Apartment	1010.0
27	42.0	Male	35123.0	Married	2.0	Bachelor's	Self-Employed	31.39190719557820	Suburban	Premium	0.0	3.0	486.0	4.0	2023-09-03 15:21:39.237118	Good	Yes	Rarely	Condo	2680.0
28	43.0	Male	1060.0	Married		Master's	Self-Employed	41.37671628604050	Suburban	Premium	0.0	8.0	795.0	7.0	2019-09-06 15:21:39.171102		Yes	Rarely	Condo	699.0
29	64.0	Male	131038.0	Single		PhD	Self-Employed	20.71707646913310	Rural	Premium		14.0	495.0	1.0	2020-08-24 15:21:39.155231	Good	Yes	Monthly	House	2360.0
30	52.0	Male	30350.0	Single	1.0	Bachelor's	Employed	21.49661354547670	Rural	Comprehensive		18.0	449.0	1.0	2023-12-28 15:21:39.192212	Poor	No	Rarely	House	641.0
31	37.0	Male	4534.0	Divorced		Bachelor's		26.387480918831300	Suburban	Premium	0.0	1.0		1.0	2022-09-21 15:21:39.102694	Poor	No	Daily	House	165.0
32	58.0	Male	1288.0	Divorced	4.0	Master's	Self-Employed	15.171628553438800	Urban	Basic	1.0	2.0	803.0	2.0	2023-07-25 15:21:39.147735	Good	No	Rarely	Apartment	3299.0
33	40.0	Female	1190.0	Single	2.0	High School	Employed	47.85157218465030	Rural	Basic		0.0	421.0	4.0	2020-08-23 15:21:39.228521	Good	No	Weekly	Condo	1128.0
34	25.0	Female	40887.0	Single	3.0	Bachelor's		17.283651290198900	Suburban	Premium		5.0	695.0	2.0	2023-04-01 15:21:39.086123	Average	Yes	Monthly	Apartment	2152.0
35	49.0	Female	12982.0	Single	4.0	PhD		32.75161705375050	Suburban	Basic	1.0	1.0	713.0	4.0	2024-02-25 15:21:39.133866	Average	No	Rarely	Apartment	30.0
36	41.0	Female		Married	3.0	PhD	Self-Employed	14.001630025186600	Urban	Comprehensive	0.0	15.0	589.0	6.0	2023-03-01 15:21:39.149731	Average	No	Rarely	House	250.0
37	50.0	Male	18313.0	Divorced	2.0	Master's	Employed	19.99646445133670	Suburban	Premium	0.0	5.0		6.0	2022-03-13 15:21:39.198406	Good	Yes	Weekly	Condo	1586.0
38	50.0	Female	57423.0	Single	1.0	Bachelor's	Self-Employed	14.14668411986700	Urban	Basic		6.0		3.0	2024-02-04 15:21:39.132191	Average	No	Daily	Condo	3890.0
39	40.0	Female	2858.0	Divorced	0.0	High School		38.57952270701200	Rural	Comprehensive	0.0	0.0	812.0	5.0	2021-09-18 15:21:39.176717	Poor	Yes	Daily	Condo	26.0
40	59.0	Male	1717.0	Divorced	1.0	Bachelor's	Unemployed	23.900048744874700	Urban	Basic	0.0	9.0	382.0	6.0	2021-04-22 15:21:39.226954	Good	No	Weekly	House	1813.0
41	46.0	Male	14053.0	Married	0.0	High School	Unemployed	18.56935672012440	Suburban	Comprehensive	3.0	14.0	469.0	5.0	2021-03-18 15:21:39.152732	Good	No	Daily	Condo	1211.0
42	21.0	Male	9085.0	Married	0.0	PhD	Self-Employed	25.719284477165000	Urban	Comprehensive	1.0	10.0		8.0	2023-06-18 15:21:39.201446	Poor	Yes	Weekly	Apartment	794.0
43	35.0	Male	39717.0	Divorced	1.0	PhD		7.263930529476290	Urban	Basic	0.0	10.0	530.0	2.0	2023-05-23 15:21:39.086123	Good	Yes	Daily	Apartment	551.0
44	19.0	Female	10481.0	Divorced	4.0	Master's		28.78354866016650	Rural	Premium	0.0	13.0	538.0	5.0	2024-03-04 15:21:39.134960	Poor	Yes	Weekly	Condo	1525.0
45	61.0	Male	5095.0		3.0	Master's	Employed	40.88612372940400	Suburban	Comprehensive	2.0	7.0	842.0	7.0	2020-08-04 15:21:39.269494	Poor	No	Weekly	House	23.0
46	31.0	Female	27950.0	Married	2.0	PhD		22.93686152514120	Urban	Basic	0.0	1.0	575.0	5.0	2020-04-20 15:21:39.170099	Poor	No	Weekly	House	1069.0
47	54.0	Female	29076.0	Married	1.0	High School	Employed	36.83989236833920	Rural	Comprehensive	0.0	2.0	685.0	7.0	2023-06-21 15:21:39.272456	Average	Yes	Daily	Condo	1645.0
48	34.0	Female	12221.0	Divorced	0.0	High School	Employed	37.82939868635600	Rural	Comprehensive	1.0	1.0	794.0	1.0	2020-03-02 15:21:39.237118	Average	Yes	Weekly	House	1486.0
49	45.0	Male	11584.0	Married	4.0	Bachelor's	Self-Employed	29.18189999982650	Urban	Basic	1.0	3.0	836.0	4.0	2020-12-30 15:21:39.287115	Good	No	Daily	Apartment	1309.0
50	30.0	Female	8599.0	Single	1.0	PhD	Employed	28.744667545434600	Rural	Comprehensive		18.0	784.0	7.0	2020-04-10 15:21:39.187211	Good	No	Rarely	House	882.0
51	59.0	Male	523.0	Divorced	1.0	PhD			Urban	Premium	1.0	9.0	439.0	9.0	2021-01-02 15:21:39.248098	Average	Yes	Weekly	Apartment	537.0
52	33.0	Male	54962.0	Married	0.0	Master's	Employed	20.08451192668930	Urban	Premium		9.0	381.0	9.0	2020-01-19 15:21:39.237118	Average	No	Daily	House	1358.0
53	50.0	Female	4928.0	Divorced	1.0	PhD	Unemployed	17.8827123747686	Suburban	Comprehensive		3.0	801.0	7.0	2020-10-19 15:21:39.280784	Good	Yes	Weekly	Apartment	2312.0
54	43.0	Female	2361.0	Single	2.0	PhD	Employed	24.39805204006730	Suburban	Comprehensive	1.0	16.0		6.0	2019-11-28 15:21:39.206847	Average	No	Daily	Condo	2093.0
55	28.0	Male	2476.0	Married	3.0	Bachelor's		19.81210975999020	Rural	Comprehensive	1.0	1.0	624.0	4.0	2019-11-16 15:21:39.194829	Average	No	Rarely	Apartment	1595.0
56	52.0	Female	27128.0	Married	1.0	Master's		35.09675695651840	Rural	Premium	3.0	0.0	439.0	2.0	2023-11-11 15:21:39.191219	Poor	Yes	Weekly	House	798.0
57	58.0	Male	78986.0	Divorced	2.0	PhD	Self-Employed	22.692009207964300	Urban	Premium		16.0	497.0	9.0	2020-11-18 15:21:39.155231	Good	No	Monthly	Apartment	78.0

データから読み取れること

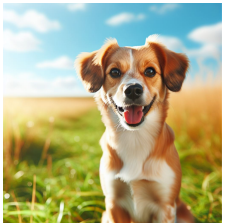
- データ数多すぎる (train120 万, test80 万)
  - 処理重いかも？バリデーションの方法も要検討。
- 数値、カテゴリ、時間が混在 & 欠損多い & 歪んだ分布
  - 前処理が絶対面倒。各カラムの特性に応じた処理を要する。

Contents

ニューラルネットワークの概要

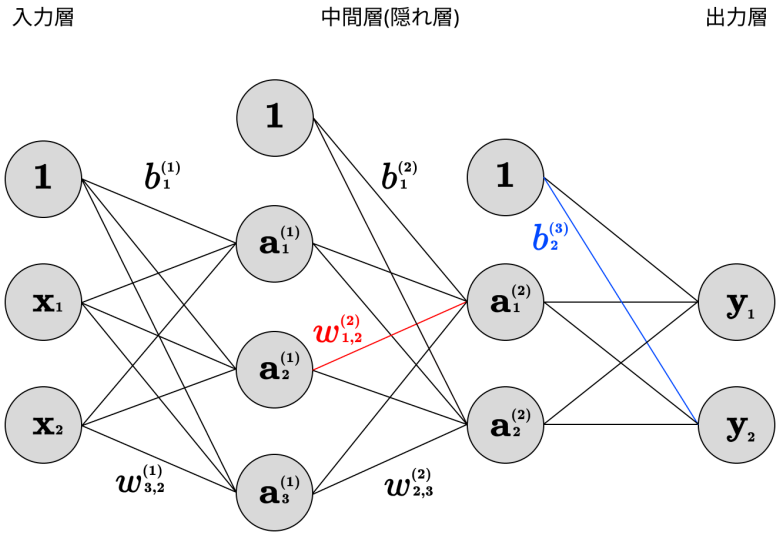
ニューラルネットワーク (NN) とは

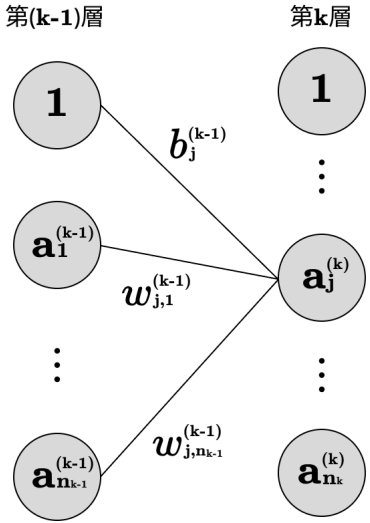
脳の神経細胞（ニューロン）のネットワーク構造を模倣し、データからパターンを学習するためのモデル。例えば、画像認識や音声認識のタスクでは、NN に大量の画像や音声のデータを与えると、層ごとに特徴を捉えて、最終的に「これは犬の画像」「これは人間の声」といった判別・分類ができるようになる。欠損値は扱えない。回帰、分類タスクいずれも利用可能。CNN, RNN, DNN, BNN など、さまざまな派生型が存在し、入力データの特性に応じて使い分けられるが、今回は標準の NN を使用する。



なぜニューラルネットワークか

- もともとニューラルネットワークに興味があった。
- 昨年度の B4 統計コンペで GBDT を用いていたため。
- オリジナリティが発揮できる。(GBDT では前処理や層の設計などをあまり要さない)


入力層
中間層(隠れ層)
出力層
1
1
x_1
x_2
$b_1^{(1)}$
$w_{3,2}^{(1)}$
1
$a_1^{(1)}$
$a_2^{(1)}$
$a_3^{(1)}$
$w_{1,2}^{(2)}$
$b_1^{(2)}$
$w_{2,3}^{(2)}$
1
$a_1^{(2)}$
$a_2^{(2)}$
$b_2^{(3)}$
y_1
y_2



$$a_j^{(k)} = h\left(\sum_{i=1}^{n_{k-1}} w_{j,i}^{(k)} a_i^{(k)} + b_j^{(k)}\right) \tag{1}$$

h : **活性化関数** (NN に**非線形性をもたらす**)
ex. ReLU(Rectified Linear Unit)

$$ReLU(u) := max(u, 0) \tag{2}$$

行列とベクトルによる表記

$$\boldsymbol{a}^{(k)} = h(W^{(k)}\boldsymbol{a}^{(k-1)} + \boldsymbol{b}^{(k)}) \tag{3}$$

$$= h\left(\begin{pmatrix} w_{1,1}^{(k)} & w_{1,2}^{(k)} & \dots & w_{1,n_{k-1}}^{(k)} \\ w_{2,1}^{(k)} & w_{2,2}^{(k)} & \dots & w_{2,n_{k-1}}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n_k,1}^{(k)} & w_{n_k,2}^{(k)} & \dots & w_{n_k,n_{k-1}}^{(k)} \end{pmatrix} \begin{pmatrix} a_1^{(k-1)} \\ a_2^{(k-1)} \\ \vdots \\ a_{n_{k-1}}^{(k-1)} \end{pmatrix} + \begin{pmatrix} b_1^{(k)} \\ b_2^{(k)} \\ \vdots \\ b_{n_{k-1}}^{(k)} \end{pmatrix}\right) \tag{4}$$

$$= \begin{pmatrix} h(w_{1,1}^{(k)} a_1^{(k-1)} + w_{1,2}^{(k)} a_2^{(k-1)} + \ldots + w_{1,n_{k-1}}^{(k)} a_{n_{k-1}}^{(k-1)} + b_1^{(k)}) \\ h(w_{2,1}^{(k)} a_1^{(k-1)} + w_{2,2}^{(k)} a_2^{(k-1)} + \ldots + w_{2,n_{k-1}}^{(k)} a_{n_{k-1}}^{(k-1)} + b_2^{(k)}) \\ \vdots \\ h(w_{n_k,1}^{(k)} a_1^{(k-1)} + w_{n_k,2}^{(k)} a_2^{(k-1)} + \ldots + w_{n_k,n_{k-1}}^{(k)} a_{n_{k-1}}^{(k-1)} + b_{n_{k-1}}^{(k)}) \end{pmatrix} \tag{5}$$

損失関数 (Loss Function)

NN の出力した値と、教師データのラベルの差異を示す関数。
回帰タスクではしばしば二乗誤差関数を用いる。

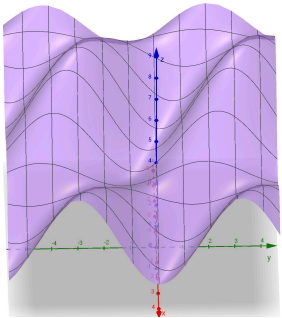
$$C(\boldsymbol{y}, \boldsymbol{t}) = \frac{1}{2} \sum_{i=1}^{n_L} (y_i - t_i)^2 \tag{6}$$

NN のおける「学習」とは、損失関数を最小化する重みパラメータを探すことである。しかし、パラメータ数が膨大な NN では、偏微分して $= 0$ の連立方程式を解くのは不可能。

勾配降下法

そこで、以下のように、損失関数が減少する方向に重みパラメータを更新する。ただし、$\eta \in [0, 1]$ は学習率とする。

$$w_{j,i}^{(l)} \leftarrow w_{j,i}^{(l)} - \eta \frac{\partial C}{\partial w_{j,i}^{(l)}} \tag{7}$$



Contents

欠損値処理

- 欠損が少ない変数
  - Age, Annual Income, Vehicle Age, Insurance Duration は平均値補完
  - Marital Status は最頻値補完
- 欠損が多い変数は、欠損であるという情報を反映
  - Number of Dependents, Occupation, Health Score, Previous Claims, Credit Score, Customer Feedback は-1 で補完

```
 #   Column                Non-Null Count    Dtype
---  ------                --------------    -----
 0   Age                   1181295 non-null  float64
 1   Gender                1200000 non-null  object
 2   Annual Income         1155051 non-null  float64
 3   Marital Status        1181471 non-null  object
 4   Number of Dependents  1090328 non-null  float64
 5   Education Level       1200000 non-null  object
 6   Occupation            841925 non-null   object
 7   Health Score          1125924 non-null  float64
 8   Location              1200000 non-null  object
 9   Policy Type           1200000 non-null  object
 10  Previous Claims       835971 non-null   float64
 11  Vehicle Age           1199994 non-null  float64
 12  Credit Score          1062118 non-null  float64
 13  Insurance Duration    1199999 non-null  float64
 14  Policy Start Date     1200000 non-null  object
 15  Customer Feedback     1122176 non-null  object
 16  Smoking Status        1200000 non-null  object
 17  Exercise Frequency    1200000 non-null  object
 18  Property Type         1200000 non-null  object
 19  Premium Amount        1200000 non-null  float64
dtypes: float64(9), object(11)
```

外れ値処理

Annual Income と Previous Claims は外れ値を持つので、IQR(Inter-Quantile Range: 四分位範囲) 法を用いる。

IQR 法

Q_1, Q_3 をそれぞれ分布の上側 25% 点、75% 点とする。データの下限, 上限 L, U を

$$IQR = Q_3 - Q_1 \tag{8}$$
$$L = Q_1 - 1.5 \times IQR \tag{9}$$
$$U = Q_3 + 1.5 \times IQR \tag{10}$$

とし、下限未満の値を下限に、上限を超える値を上限に丸め込む

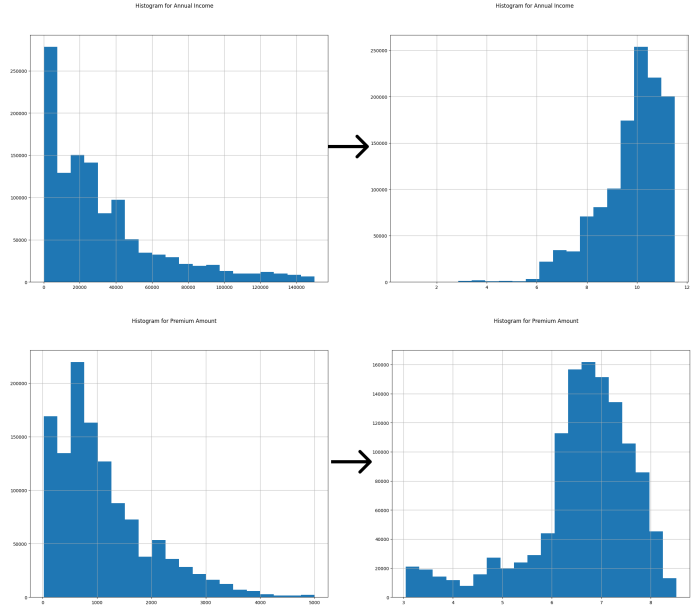
歪んだ分布

Annual Income と Premium Amount は分布が偏っているので、

対数変換

$$X \leftarrow \log(X+1) \tag{11}$$

(1 を足すのは、小さい値で発散 (オーバフロー) するのを防ぐため)
目的変数である Premium Amount については、出力する際に逆変換 $(e^X - 1)$ をして出力する。


Histogram for Annual Income
Histogram for Annual Income
Histogram for Premium Amount
Histogram for Premium Amount

時間データ処理

時間データはそのまま処理すると、例えば 2024-12-31 の次が 2025-1-1 というつながりを処理できないので、周期性や循環性をもたせる。

三角関数変換

$$Year_{sin} = \sin\left(2\pi \times \frac{Year - Year_{min}}{Year_{max} - Year_{min}}\right) \tag{12}$$

$$Year_{cos} = \cos\left(2\pi \times \frac{Year - Year_{min}}{Year_{max} - Year_{min}}\right) \tag{13}$$

Month, Day についても同様。

数値変数のスケーリング

標準化スケーリング

μ : 平均, σ : 標準偏差 とし、

$$X \leftarrow \frac{X - \mu}{\sigma} \tag{14}$$

$$\tag{15}$$

と変換する。

新たな特徴量の作成

保険利用年数と車両の乗車年数の積として、新たな特徴量
Duration mul Vehicle Age = Insurance Duration × Vehicle Age
を作成した。

カテゴリ変数のエンコーディング

カテゴリ変数 (Gender, Marital Status, Education Level, Occupation, Location, Policy Type, Customer Feedback, Smoking Status, Exercise Frequency, Property Type) **について、**One-Hot **エンコーディングを施す。例えば、**Gender **について**

	Gender_Female	Gender_Male
Female	1	0
Male	0	1

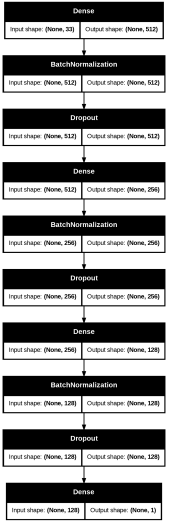
ただし、多重共線性が生じないように、カテゴリ変数の水準 c **個あるとき、**$c-1$ **個のダミー変数を作成する。**

Contents

モデルの訓練

train データの一部を、バリデーションに使用する。データ数は十分にあるので 5%(6 万件) をバリデーションに用いる。データ数が多いので、クロスバリデーションはしない。

NN の構成

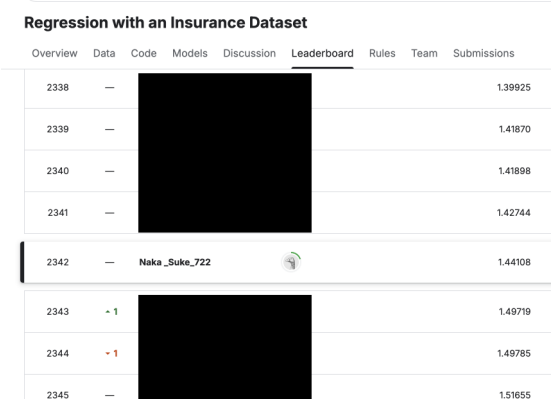

Dense
Input shape: (None, 33)
Output shape: (None, 512)
BatchNormalization
Input shape: (None, 512)
Output shape: (None, 512)
Dropout
Input shape: (None, 512)
Output shape: (None, 512)
Dense
Input shape: (None, 512)
Output shape: (None, 256)
BatchNormalization
Input shape: (None, 256)
Output shape: (None, 256)
Dropout
Input shape: (None, 256)
Output shape: (None, 256)
Dense
Input shape: (None, 256)
Output shape: (None, 128)
BatchNormalization
Input shape: (None, 128)
Output shape: (None, 128)
Dropout
Input shape: (None, 128)
Output shape: (None, 128)
Dense
Input shape: (None, 128)
Output shape: (None, 1)

- Dense: 全結合
  - 隣接する層どうしのニューロンはすべて結合
- BatchNormalization: バッチ正規化
  - 学習の収束を早め、安定化
- Dropout: ランダムにニューロンを失効
  - 特定のニューロンに依存させず、過学習を防止させ、計算コストも削減

活性化関数はいずれも $ReLU$ を使用。
中間層のニューロンの数が減少 (512 → 256 → 128) しているのは、層を経るごとに情報が縮約していくため。(生物の視覚野もそのようになっている)
ニューロン 2 のべき乗としているのは、メモリアクセスと GPU の並列処理向上のため。

結果

RMSLE = 1.44108 (1 位 1.01706, 100 位 1.02965, 1000 位 1.05976, 2000 位 1.15885)
LB 順位: 2342 / 2512 位



反省点と改善策

- 前処理に手こずり、モデルに対する厳密な設計をする時間が不十分だった
  - ハイパーパラメータを理論的にチューニングしたい
- もう少しハイペースで Submit を繰り返せばよかった
  - 少しの改善でも Submit するべきだった (長考するよりも「ものは試し」)
- さまざま機械学習手法や、アンサンブルも今後実践
  - NN の派生や、GBDT などの NN 以外の手法やそれらのアンサンブルを実践 (特にテーブルデータの分析において、GBDT が有効)
  - 予測結果の不確実性など、NN の欠点を補う方法→卒論に応用へ
- プログラミングスキルの向上
- ドメイン知識を増やすべき
- Discussion(参加者による掲示板) ももう少し参考にすればよかった
- 再度別の Kaggle コンペ (可能ならチーム参加) したい

参考文献

[1] Sebastian Raschka, Vahid Mirjalili (2020). [第 3 版]Python 機械学習プログラミング 達人データサイエンティストによる理論と実践 (原題: Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn and TesnorFlow 2, 2rd Edition). 株式会社インプレス. (訳: 株式会社クイープ).
[2] 斎藤康毅 (2016). ゼロから作る Deep Learning —Python で学ぶディープラーニングの理論と実装. オライリー・ジャパン.
[3] 門脇大輔, 阪田隆司, 保坂桂佑, 平松雄司 (2019). Kaggle で勝つデータ分析の技術. 技術評論社.