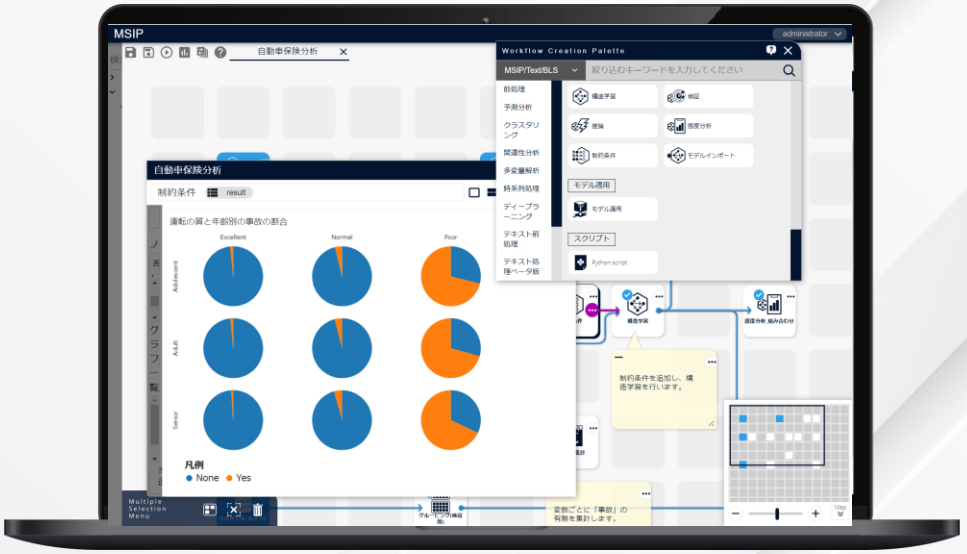
BayoLinkS

ベイジアンネットワーク構築支援システム

テクニカルサンプルプロジェクト

自動車保険データを用いた事故の要因分析



株式会社 NTTデータ数理システム

このプロジェクト について

こんな方におすすめします

カテゴリカルデータを用いた要因分析を効率的に行いたい方

BayoLinkS に搭載している各種機能の利用方法を知りたい方

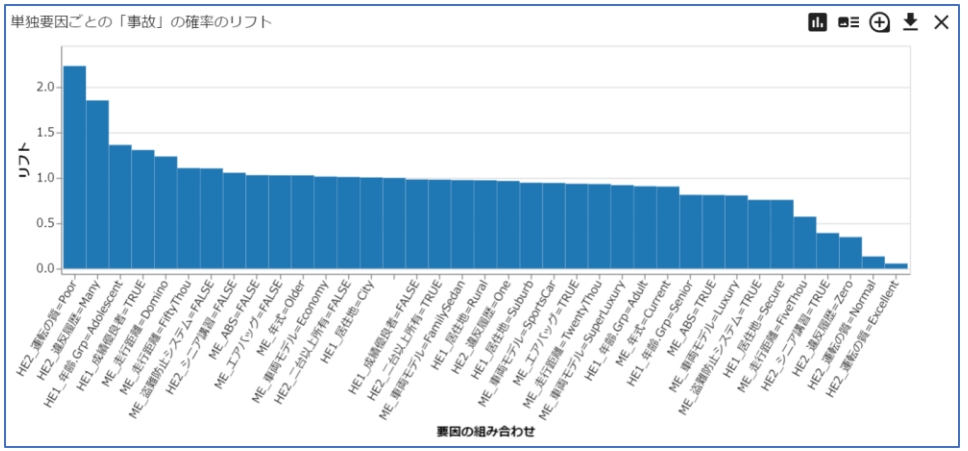
何をするプロジェクト？

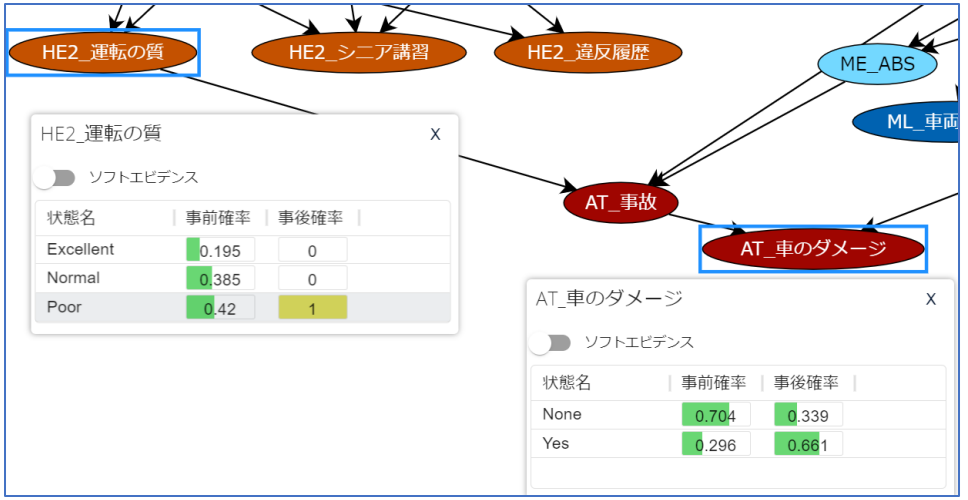
多変数かつ多カテゴリのデータを用いて、ある事象の要因を探るとき、次のような問題を抱えることが多いです。

・ **組み合わせの数が膨大**なため、要因の調査に莫大な時間を要してしまう。

・ 各変数間の関連性や因果構造が把握しづらい。

本プロジェクトでは BayoLinkS を用いて、**因果構造を把握し、要因分析を行います。**具体的には、p.4 で紹介する自動車保険データを用いて「事故」が起こる要因を探ります。





※p.25の「本文書・プロジェクトファイルのご利用にあたって」をお読みのうえ、ご利用ください。

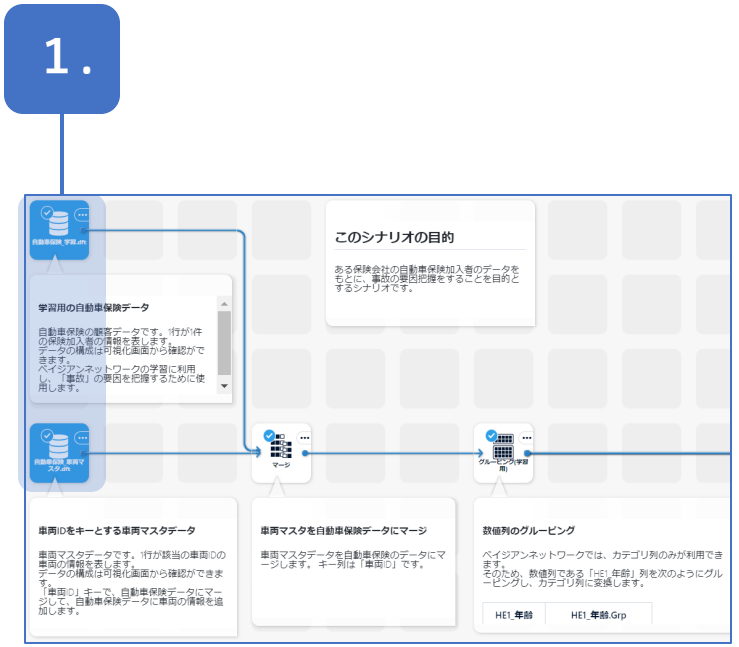
プロジェクト 解説

プロジェクト 解説

1. 対象データ

・a. 顧客データ：自動車保険の加入者の情報や事故の履歴を記載

・b. 車両マスタ：各車両ごとの特徴を記載



自動車保険のデータ (一部抜粋)

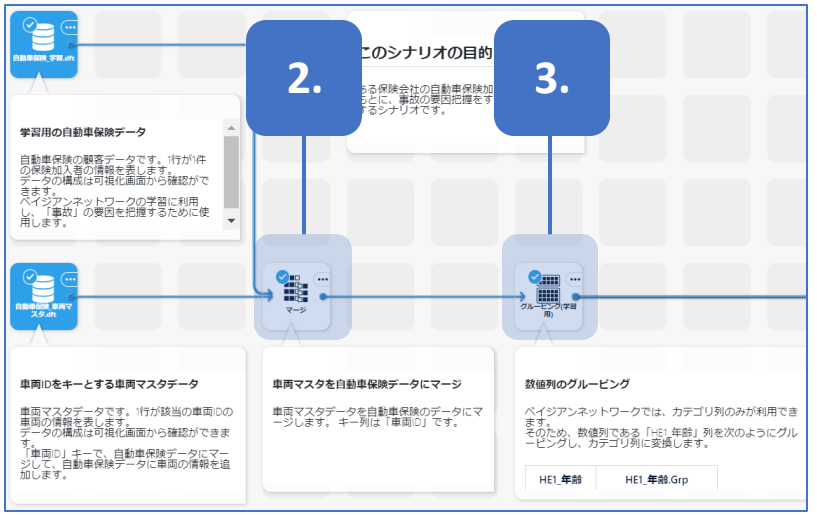
顧客データ	
項目	値
年齢	数値データ
居住地	City/Suburb/Secure/Rural
運転の質	Excellent/Normal/Poor
違反履歴	Many/One/Zero
車両ID	識別番号
事故	Yes/None
車のダメージ	Yes/None

車両マスタ	
項目	値
車両ＩＤ	識別番号
年式	Current/Older
車両モデル	Luxury/Economy/Super Luxury
ABS	TRUE/FALSE
エアバック	TRUE/FALSE
盗難防止システム	TRUE/FALSE

プロジェクト 解説

2. マージ

顧客データと車両マスタをマージします。
キー列は、「車両ID」です。



3. 「年齢」列のグルーピング

グルーピングノードを用いて、数値データである年齢列をグルーピングします。若年層や高齢者は事故を起こしやすいことが予想されるため、若年層、青年、高齢者の3つのカテゴリでグルーピングを行います。

年齢	カテゴリ名
0-19	Adolescent（若年層）
20-59	Adult（青年）
60-	Senior（高齢者）

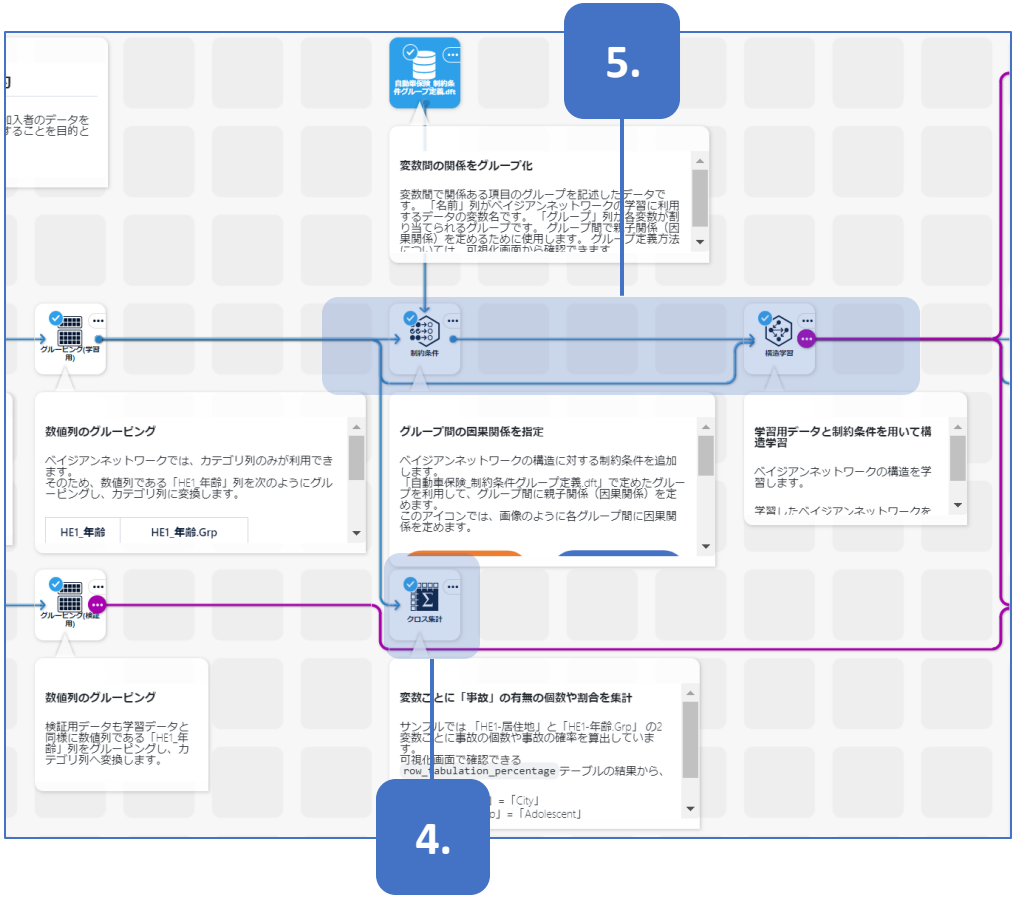
※ベイジアンネットワークを構築するためには、数値型の列は、カテゴリ型にする必要があります。カテゴリ件数は、5件以下が扱いやすいため、今回は3件のカテゴリに分けています。

4. クロス集計

各列の値ごとにある列の値の個数や割合を集計します。
プロジェクトでは、「居住地」と「年齢」ごとの「事故」の割合を集計しています。

5. 構造学習

ベイジアンネットワークの構造をデータから学習する**構造学習**を行います。
「制約条件」で**ベイジアンネットワークの親ノード（原因）の候補となる変数の集合を指定します。**このようにデータからは学習出来ない情報を外部から与えることで、より現実に即した因果関係を表すベイジアンネットワークモデルを作成することができます。



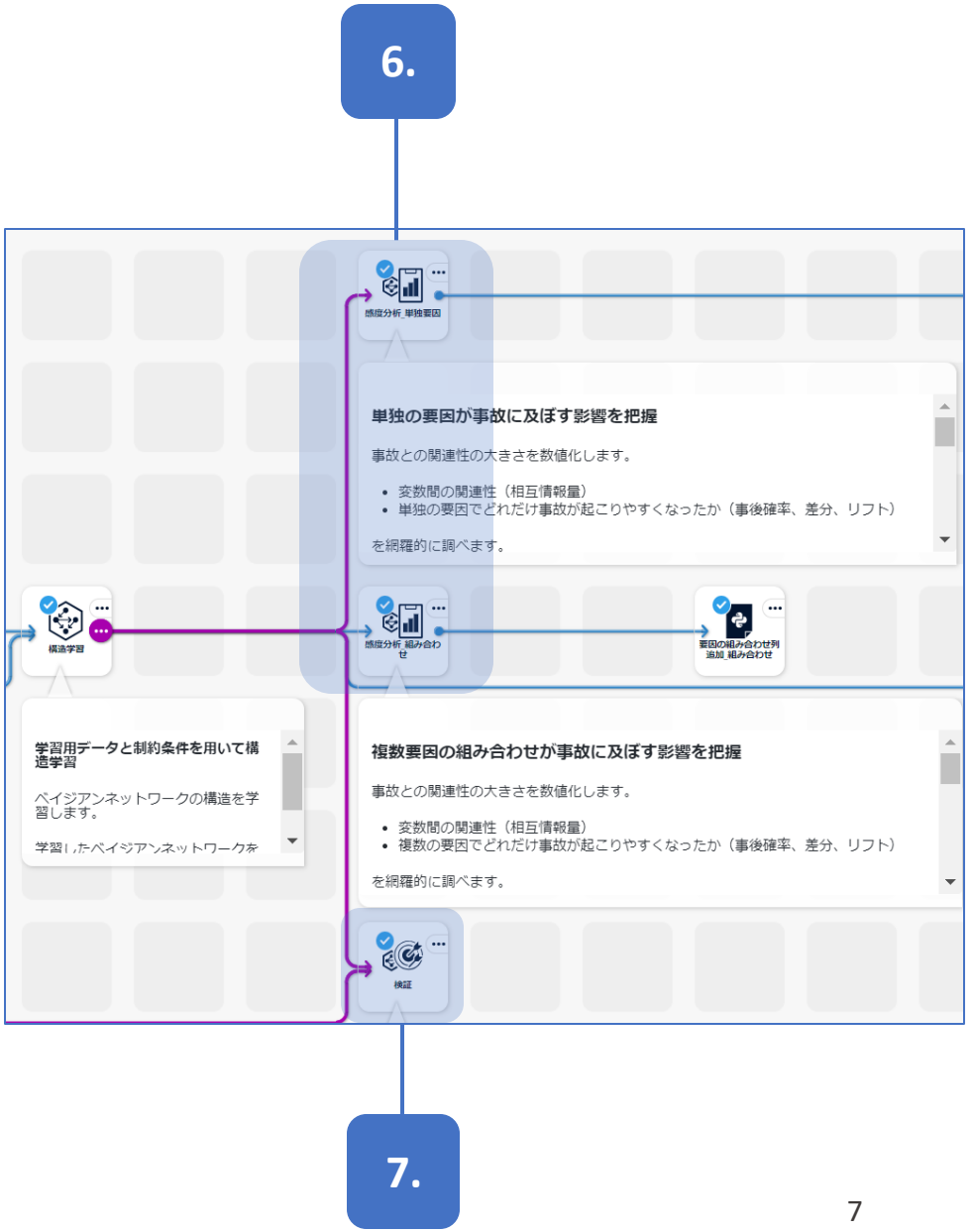
プロジェクト 解説

6. 感度分析

変数と「事故」の**関連性の大きさを数値化**します。また、エビデンスを追加したときの確率値の差分を算出します。
「事故」を起こす要因を把握し、自動車保険のサービス内容の考案やサービス内容の見直しに役立てます。

7. 検証

未知のデータに対して学習データと同じ前処理を行ったうえで、学習済みモデルによる予測を行います。
ベイジアンネットワークモデルの**予測精度の把握**し、ベイジアンネットワークの妥当性を検証します。

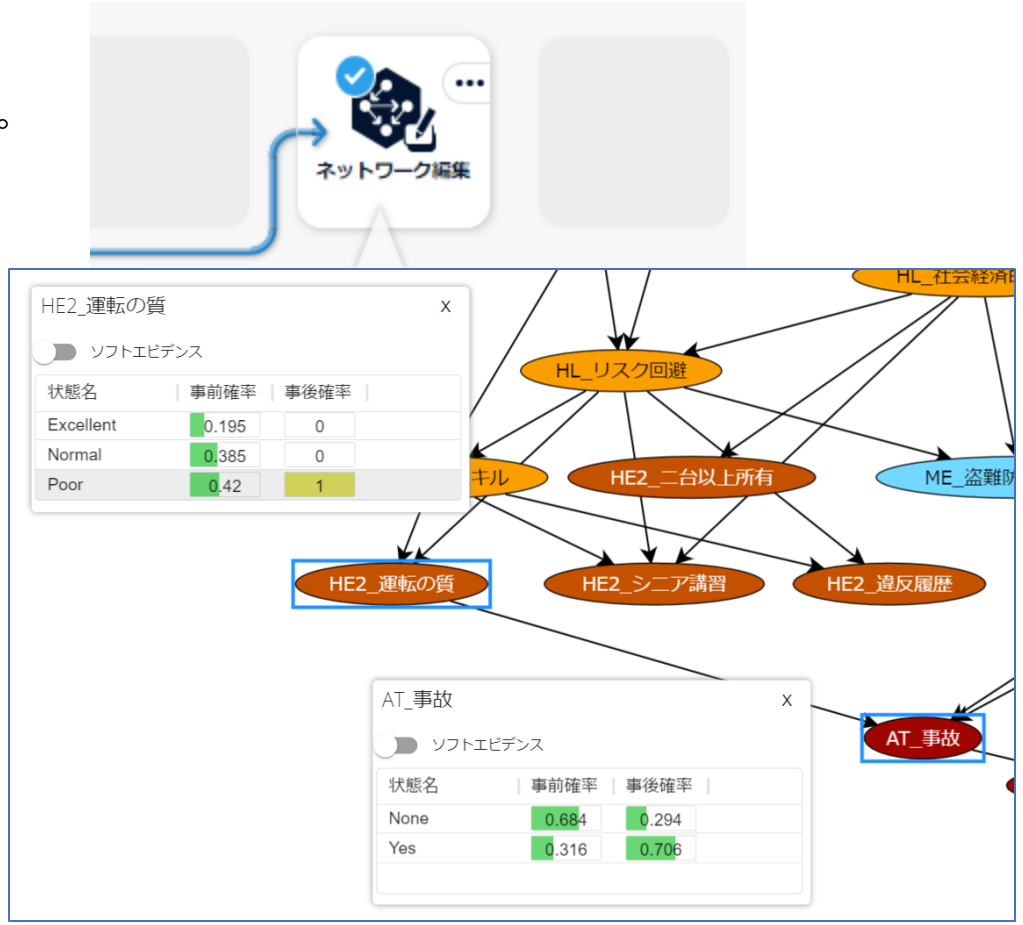


8. ネットワーク編集

学習したベイジアンネットワークモデルの可視化をします。
ネットワーク構造を視覚的に見ることで、因果関係の直感的な把握に役立ちます。

ベイジアンネットワーク上で**確率推論**を行います。
確率推論は、「ある変数を変化させる要因となる値は何か？」を探索的に調べるときに利用します。

例えば、「運転の質が悪い人は、通常と比べてどれだけ事故の可能性を上げるのか？」といった内容を確率的に計算することができます。

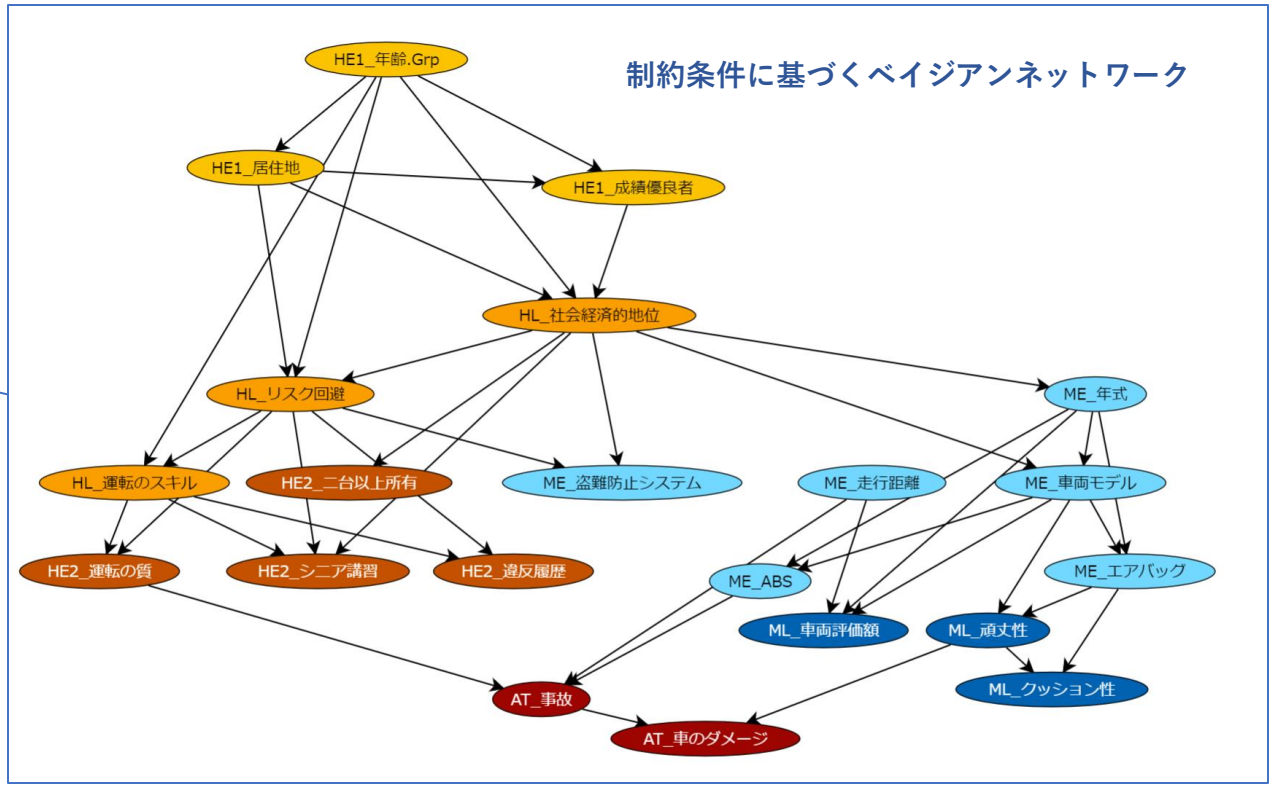


アウトプットの説明

アウトプット

ベイジアンネットワークモデル

本プロジェクトでは、制約条件に基づき生成したベイジアンネットワークモデルを出力します。変数間の因果関係の把握に役立ちます。

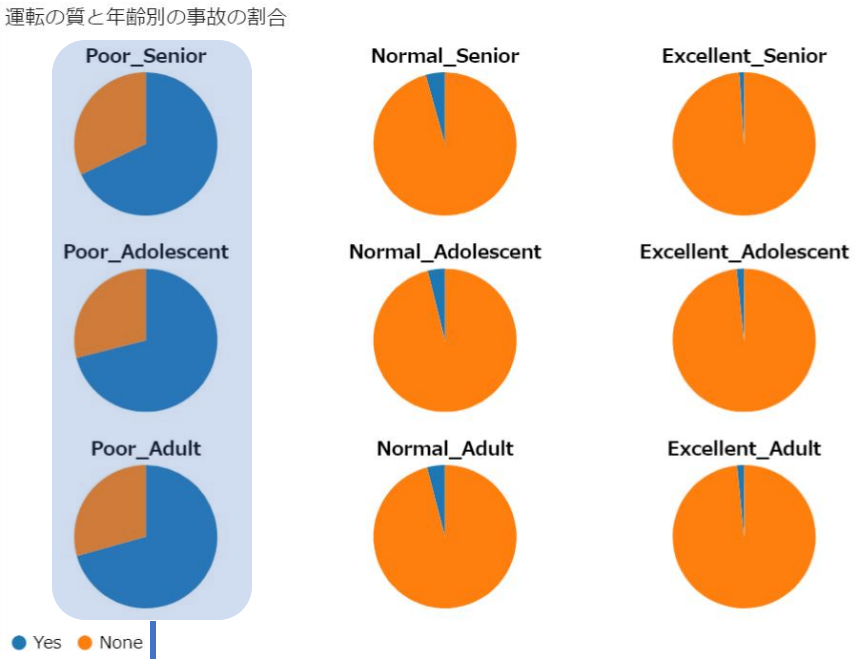


アウトプット

自動車保険データの各種プロット

可視化画面から、次のプロットが確認できます

1. **各カテゴリ値ごとの事故の割合**
   事故の割合がカテゴリごとにどの程度変わるかの直観的な把握に利用します
2. **「事故」との相互情報量**
   事故に影響が大きい変数の把握に利用します
3. **エビデンスごとの「事故」の確率の差分**
   事故に影響が大きい変数と値の把握に利用します



運転の質によって事故割合が大きく異なり、Poorの場合は特に事故割合が高い

アウトプット

感度分析

1. 感度分析_単独要因

1変数ごとに事故に影響を及ぼす
大きさを出力します
事故の要因把握に利用します

2. 感度分析_組み合わせ

複数の変数の組合せが事故に与える
影響の大きさを出力します
複数の変数の組み合わせを見ることで
事故の要因把握をより詳細に
行うことが出来ます

感度分析_単独要因-high　列数: 19 行数: 74

No.	目的変数 Category	値 Cate	事前確率 Float	事後確率 Float	差分 Float	リフト Float
1	AT_事故	Yes	0.316056	0.706361	0.390304	2.234919
2	AT_事故	Yes	0.316056	0.586323	0.270266	1.855120
3	AT_事故	Yes	0.316056	0.431087	0.115030	1.363954

出力のhighテーブルでは、
事前確率（要因追加前の確率）、事後確率（要因追加後の確率）とその差分、リフトなど
要因の大きさを測る指標が表示されます

リフト Float	HE1_成績優良者 Category	HE1_居住地 Category	HE2_運転の質 Category	HE2_シニア講習 Category	HE2_ Ca
2.234919			Poor		
1.855120					Many
1.363954					

リフトより右の列では確率値を変化させる
要因（運転の質=Poorなど）が表示されます

アウトプット

検証

構築したモデルの予測精度を確認します。
ベイジアンネットワークの妥当性を検証します。
出力は次の3つのテーブルです。

- result テーブル
- summary テーブル
- infer テーブル

result テーブル、summary テーブルからは次のことが分かります。

- AT_事故の正解率：0.878（事故の有無の予測を87.8%当てている）
- AT_事故=Yes の適合率：0.750（事故を起こすと予測した保険加入者のうち、75%が実際に事故を起こしている）
- AT_事故=Yes の再現率：0.876（実際に事故を起こした保険加入者のうち、87.6%が事故を起こすと予測している）

このことから、今回構築したモデルの妥当性は十分に高いと言えます。

また、infer テーブルでは、保険加入者ごとの事故を起こす確率値の予測値を確認できます。

構造学習で構築する
自動車保険モデルの解説

自動車保険モデルの因果関係の仮説

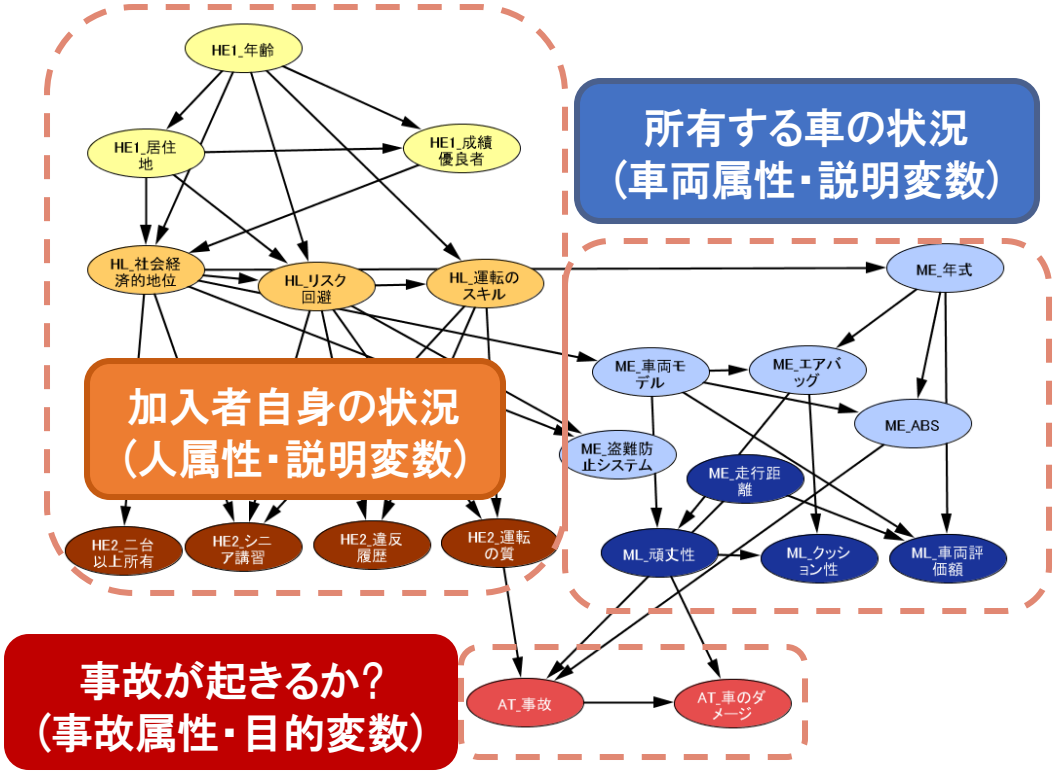
変数を属性ごとにグループ化する

どのような因果仮説のもと自動車保険モデルに
制約条件を与えたかについて解説します。

因果関係の整理のため、与えらえたデータの変数を
属性ごとにグループ化します。

「事故の発生は、人属性と車両属性に依存する」
と考え、学習データの変数を次の3属性にグループ化
します。

- **人属性**
- **車両属性**
- **事故属性**



※本サンプルプロジェクトでは、制約条件を与えてベイジアンネットワークを作成しています。
これはデータからは学習出来ない因果構造を制約条件としてネットワーク構造に追加することで
要因把握に有益なベイジアンネットワークを作成することが出来るためです。

自動車保険モデルの因果関係の仮説

因果関係の設定

因果関係の整理のため、各属性内の変数をさらに細かくグループに分割します

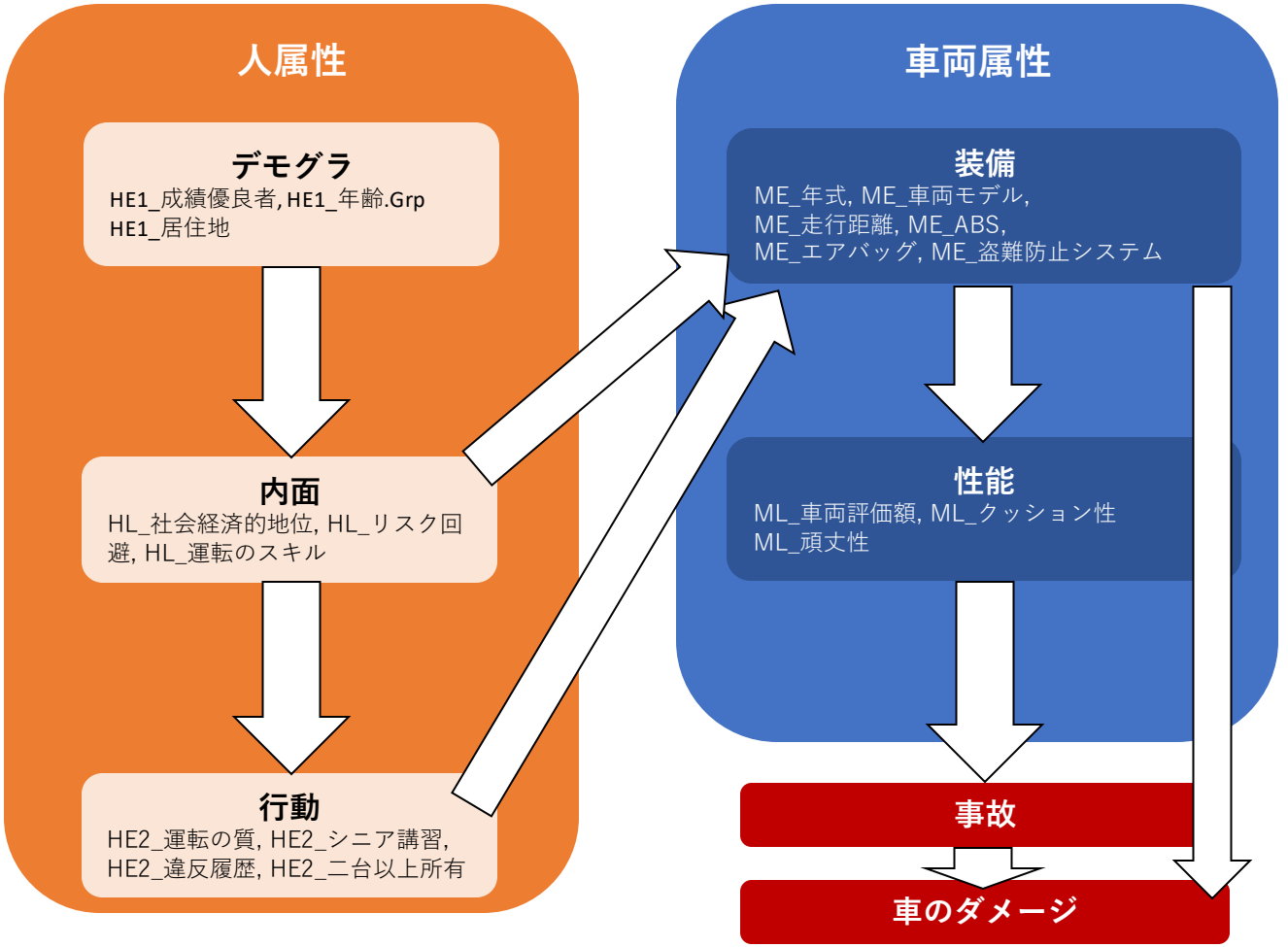
人属性を次の3つのグループに分割

- **デモグラ**
- **内面**
- **行動**

車両属性を次の2つのグループに分割

- **装備**
- **性能**

グループ間の因果関係を図のように矢印で繋ぎます。



制約条件の設定方法

1. 制約条件グループを記述したファイルの作成

制約条件グループを記述したファイルを作成します。1列目に変数名、2列目に変数が所属するグループ名を記載した制約条件ファイルを作成し、MSIP にインポートします。

No.	名前 CATEGORY	グループ CATEGORY
1	HE1_成績優良者	人属性(デモグラ)
2	HE1_年齢.Grp	人属性(デモグラ)
3	HE1_居住地	人属性(デモグラ)
4	HL_社会経済的地位	人属性(内面)
5	HL_リスク回避	人属性(内面)
6	HL_運転のスキル	人属性(内面)

2. 「制約条件」の設定

「制約条件」の「グループ設定」から、親ノード候補となるグループを設定します。
「制約条件」から出力される制約条件設定用のデータを用いて構造学習を行います。

グループ設定

グループ名	列名	親候補
人属性(デモグ	HE1_成績優良	人属性(デモグラ)
人属性(内面)	HL_社会経済的	人属性(デモグラ),人属性(内面)
人属性(行動)	HE2_運転の質	人属性(内面),人属性(行動)
車両属性(装備	ME_年式,ME_	人属性(内面),人属性(行動),車両属性(装備)
車両属性(性能	ML_車両評価	車両属性(装備),車両属性(性能)
車のダメージ	AT_車のダメ	車両属性(性能),事故
事故	AT_事故	人属性(行動),車両属性(装備)

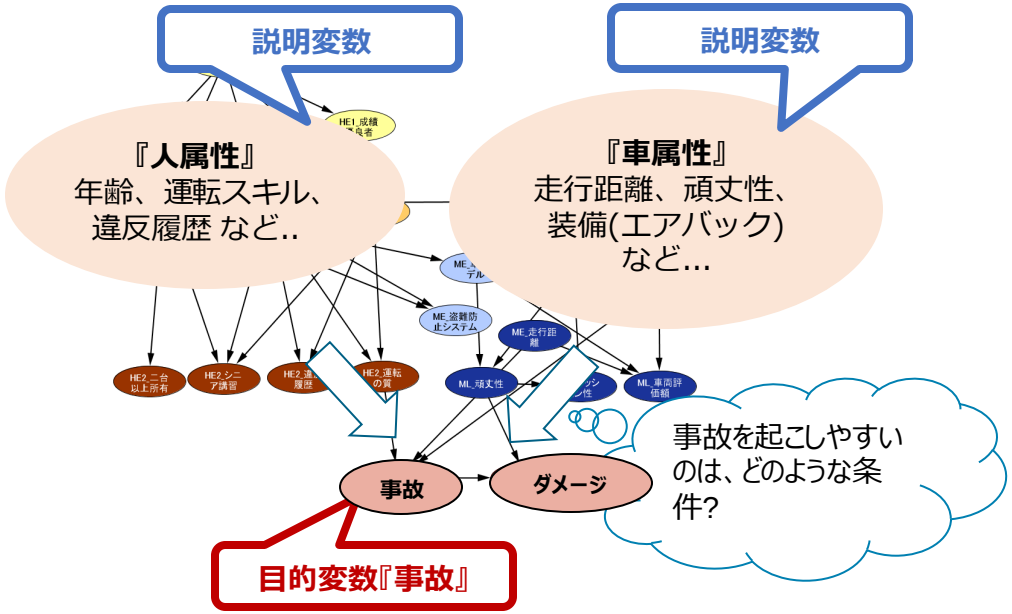
感度分析による要因分析の解説

感度分析

感度分析とは

感度分析は、説明変数のうちどの変数が目的変数により大きく影響を与えているか、を探る分析です。

本プロジェクトでは、自動車保険モデルで、**事故の発生確率にはどのような要因が寄与するか**を調べています。



自動車保険モデルの感度分析

単独要因での事故への影響を評価

「感度分析_単独要因」で、各説明変数の目的変数（本プロジェクトでは「事故」）への影響を調べます。

出力テーブルの説明

・mi

「事故」と関連性を**相互情報量**という指標で出力します。相互情報量が大きいほど関連性が強いとされています。

感度分析_単独要因-mi　列数: 3 行数: 13

No.	目的変数 CATEGORY	説明変数 CATEGORY	相互情報量 FLOAT
1	AT_事故	HE2_運転の質	0.411036
2	AT_事故	HE2_違反履歴	0.173687
3	AT_事故	HE1_居住地	0.032311
4	AT_事故	HE1_年齢.Grp	0.019075
5	AT_事故	HE2_シニア講習	0.014325

・high/low

説明変数が指定される前の確率（事前確率）と後の確率（事後確率）の差分やリフト値を出力します。

感度分析_単独要因-high　列数: 19 行数: 74

No.	目的変数 CATEGORY	値 CATE	事前確率 FLOAT	事後確率 FLOAT	差分 FLOAT
1	AT_事故	Yes	0.316056	0.706361	0.390304
2	AT_事故	Yes	0.316056	0.586323	0.270266
3	AT_事故	Yes	0.316056	0.431087	0.115030
4	AT_事故	Yes	0.316056	0.413462	0.097405
5	AT_事故	Yes	0.316056	0.390943	0.074887
6	AT_事故	Yes	0.316056	0.350383	0.034327

単独要因での感度分析の結果の考察

相互情報量と確率値の差分のまとめ

	変数名	相互情報量
1	運転の質	0.41
2	違反履歴	0.17
3	居住地	0.03
4	年齢	0.02

	変数名　⇒　値	差分	リフト値
1	運転の質 ⇒ Poor	0.39	2.23
2	違反履歴 ⇒ Many	0.27	1.86
3	年齢 ⇒ Adolescent	0.12	1.36
4	成績優良者 ⇒ Yes	0.10	1.31
5	走行距離 ⇒ Domino(10万 M)	0.08	1.24

事故への影響大

- 運転の質、違反履歴、居住地、年齢　の影響が大きいことが分かります
- 事故を起こす確率を上げる要因
  - 「**運転の質が悪い**」「**違反履歴が多い**」「**若年層**」「**走行距離が長い**」
- 「成績優良者」が事故を起こす確率を上げるのは、いわゆるペーパードライバー（あまり運転をしない為、違反を起こさずに成績優良者となった人）の影響があるかもしれません

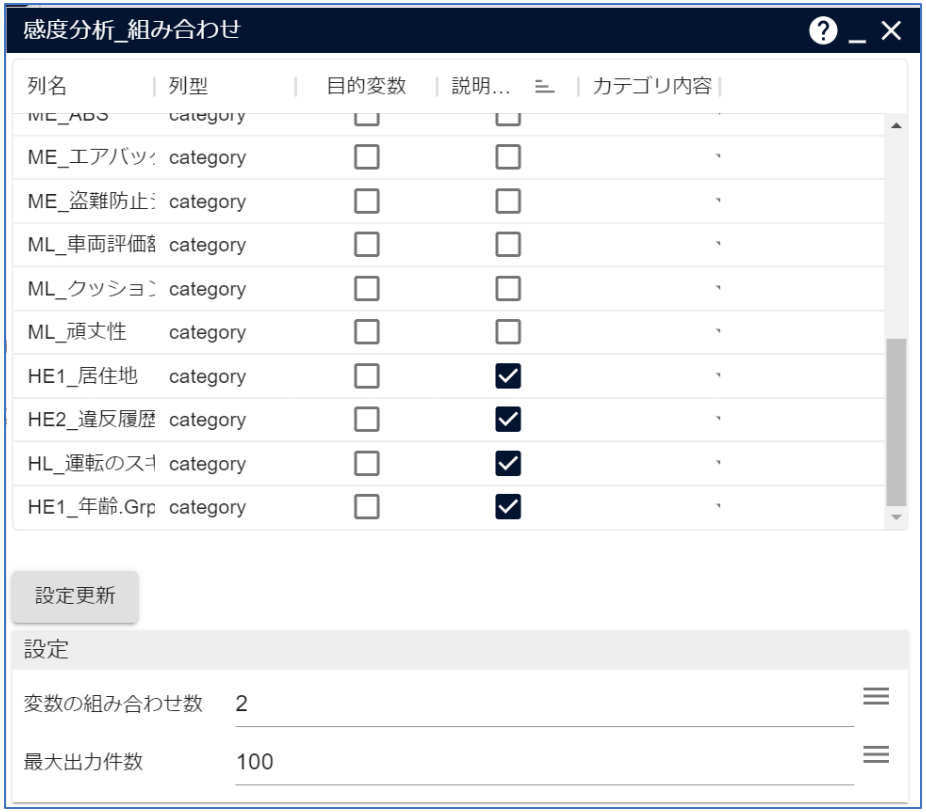
自動車保険モデルの感度分析

複数要因での事故への影響の評価

単独要因で事故への影響が大きかった「運転の質」「違反履歴」「居住地」「年齢」の変数がどのように組み合わさると事故への影響が大きくなるかを分析します。

設定項目の説明

複数要因の組み合わせの影響を見るためには、**設定項目の「変数の組み合わせ数」を増やします。**
本プロジェクトでは2となっており、2変数の組み合わせで要因の変化を見ています。



複数要因での感度分析の結果の考察

目的変数「事故」⇒ Yes のときの確率値の差分とリフト値

	変数名　⇒　値	差分	リフト値
1	運転の質 ⇒ Poor, 居住地 ⇒ City	0.3995	2.263
2	運転の質 ⇒ Poor, 居住地 ⇒ Rural	0.3936	2.245
3	運転の質 ⇒ Poor, 年齢 ⇒ Adult	0.3911	2.237
4	運転の質 ⇒ Poor, 違反履歴 ⇒ Zero	0.3908	2.237
5	運転の質 ⇒ Poor, 年齢 ⇒ Adolescent	0.3906	2.236
…	…	…	…
77	運転の質 ⇒ Excellent, 居住地 ⇒ Secure	-0.2993	0.053

- 運転の質の影響が最も大きく、それ以外の変数の影響はあまり大きくないことが分かります
- 運転の質が悪い 中で特に事故の確率を上げる要因
  - 「**都市に居住**」「**田舎に居住**」「**20-59歳の青年層**」「**19歳までの若年層**」
- 「違反履歴がなく運転の質が悪い」人が事故を起こしやすいのは、普段あまり運転をしないからである可能性があります
- 運転の質が良く、安全な地域に居住している人は事故を起こしにくい傾向にあります

補足情報

技術的な情報や利用規約について

本文書・プロジェクトファイルのご利用にあたって

本文書ならびにプロジェクトファイルは、（株）NTT データ数理システム （以下「弊社」）が開発・販売する分析プラットフォーム BayoLinkS についての情報提供として弊社が作成を行ったものです。

プロジェクトファイルは、ご利用者様の責任のもとで改変して利用することができますが、これに対するリバースエンジニアリングを禁じます。また、本文書ならびにプロジェクトファイルのご利用に際して、ご利用者様および第三者に損害が発生したとしても、弊社は責任を負わないものとします。

プロジェクトファイルは、その中に同梱されているデータを利用し、本文書内で解説している設定可能なパラメータで動作させた場合についてのみ、弊社にて動作の検証を行っております。これを超えるような状況における動作は保証いたしません。

本プロジェクトファイル は、MSIP1.8.2 および BayoLinkS 9.1.1 にて動作確認を行っております。

BayoLinkS

ベイジアンネットワーク構築支援システム

お問い合わせ: 株式会社NTTデータ数理システム　営業部
Tel: 03-3358-6681
E-mail: bayolink-info@ml.msi.co.jp
WEB: https://www.msi.co.jp/solution/bayolinks/