Суртаева Мария

Байесовские сети

Датасет

В задании использовались предложенные данные о страховании. Задействован поднабор со следующими столбцами:

- SocioEcon - социально-экономический статус;
- RiskAversion - склонность избегать риск;
- DrivQuality - качество вождения;
- Accident - тяжесть несчастного случая;
- CarValue - стоимость автомобиля;
- ThisCarCost - расходы на застрахованный автомобиль.

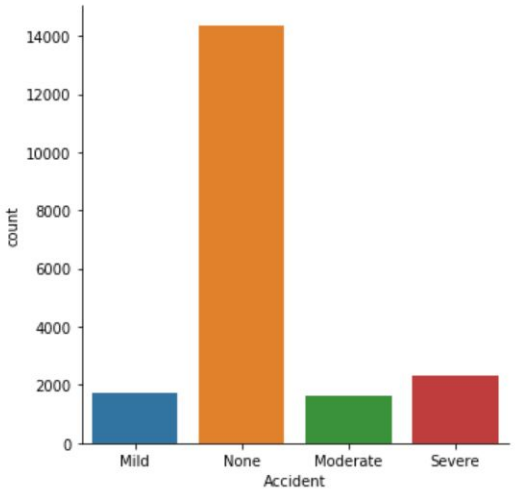


рис. 1

Все признаки категориальные, имеют 3-5 текстовых значений. В полном датасете 20000 записей без пропущенных значений и следующее распределение значений Accident (рис. 1). Выбирая между этим признаком и SocioEcon, следует отметить, что преобладание наблюдений без происшествий (None) характерно для всех групп SocioEcon (рис. 2).

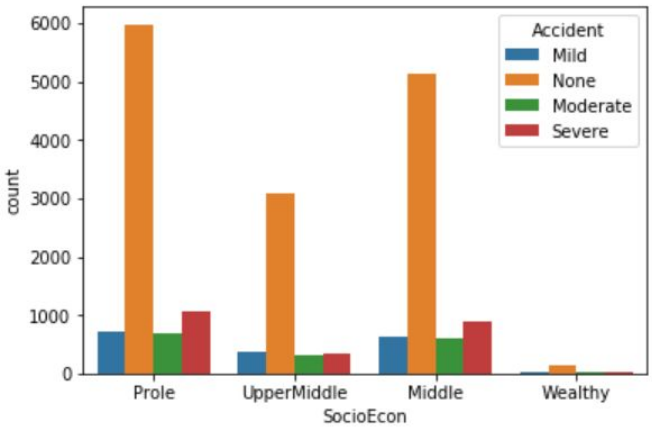


рис. 2

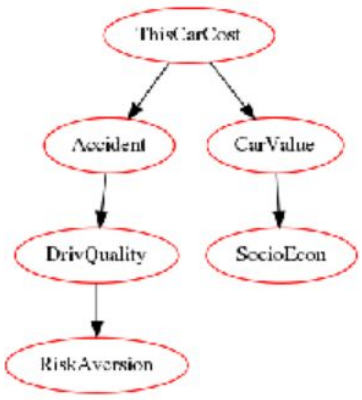
Так как интерес представляют ответы вероятностной модели в случаях происхождения несчастного случая, из датасета были отобраны только записи со значениями Moderate (умеренная тяжесть) Severe (тяжелый несчастный случай).

В итоге получено 3926 наблюдений без пропущенных значений.

Построение байесовский сетей

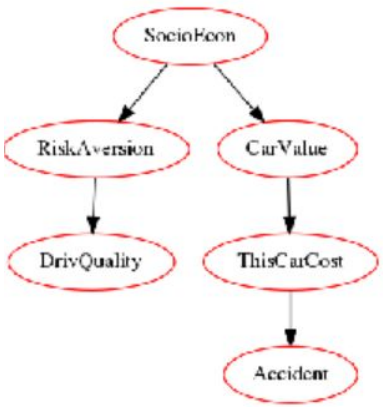
По данным были обучены 3 байесовские сети с использованием разных алгоритмов. Сети строились по обучающей выборке из датасета - 80% случайных наблюдений. Для оценки качества построенной структуры использовался логарифм правдоподобия. На тестовой выборке проводилась оценка качества предсказания моделью значения ThisCarCost с помощью точности (accuracy). Разбиение на обучение и тест было единым для всех сетей.

1. **“exact-dp”**



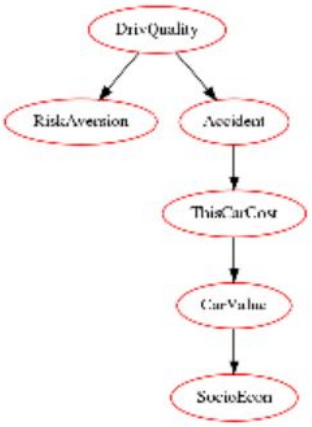
   Оценка качества структуры: -13900.

2. **“chow-liu”**



   Оценка качества структуры: -13875.

3. **“greedy”**



Оценка качества структуры: -13900.

Все сети имели одинаковое априорное распределение для расходов на застрахованный автомобиль:

```
[{'Thousand': 0.1417197452229303,
  'TenThou': 0.6181528662420374,
  'HundredThou': 0.23694267515923567,
  'Million': 0.0031847133757968364}]
```

Также предсказания моделей получили одинаковую оценку качества предсказаний: точность = 0.617.

При более детальном анализе результатов оказалось, что ответы модели содержат преимущественно один класс из четырех возможных для признака ThisCarCost, а распределение выглядит так (рис. 3). При этом и в обучающей, и в тестовой выборке содержались все четыре класса.

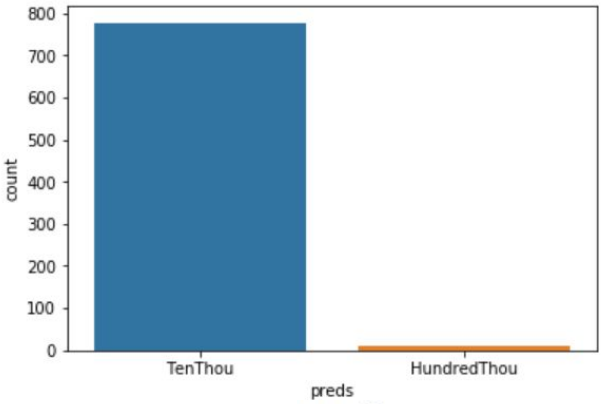


рис. 3

Заключение

Любопытным фактом данной работы служит единогласность результатов распределений и предсказаний трех по-разному обученных байесовских сетей. Несмотря на очевидное различие структур, предсказательная способность трех моделей оказалась одинакова и явно выражена за счет сильного расхождения вероятностей значений оцениваемого признака. Как и следовало ожидать,

значение с максимальной априорной вероятностью стало лидером и в частоте предсказаний.

Таким образом, итоговая модель может быть определена только основываясь на метрике качества структуры. Максимум правдоподобия сети для обучающей выборки датасета достигнут при приближенном алгоритме Chow Liu.

Выводы

Основываясь на выбранной модели, можно заключить, что социально-экономический статус водителя влияет на склонность избегать риски, и на стоимость автомобиля. Наличие тяги к риску воздействует на качество вождения, а стоимость автомобиля, в свою очередь, оказывает влияние на расходы страховки. До этого момента все звучит логично с точки зрения здравого смысла, однако на финише сеть устанавливает, что расходы на застрахованный автомобиль влияют на тяжесть несчастного случая, что сомнительно. Тем не менее, при ручном внесении ограничений на граф связей сеть осталась неизменной в своей структуре.