ZZN Projekt 2024/25

Téma 8. Databáze hodnocení aerolinek cestujícími

Úloha 2. Řešení dolovacích úloh

Tým:

Adam Hos (vedoucí) – xhosad00

Dominik Pop – xpopdo00

Obsah

1 Popis datové sady

Datová sada obsahuje údaje z průzkumu spokojenosti cestujících letecké společnosti. Shromažďuje demografické informace o cestujících, dále pak detaily o jejich letu a hodnocení různých jednotlivých vlivů před/během letu. Dataset se především zaměřuje na hodnocení služeb jako je pohodlí sedadel, kvality jídla/pití, kvality palubní obsluhy nebo třeba spolehlivosti online rezervace. Dataset se dá použít k analýze, která může letecké společnosti pomoci vylepšit své služby a porozumět lépe preferencím zákazníků.

- Id – Id záznamu
- Gender – pohlaví cestujícího (string – muž / žena)
- Customer Type – loajálnost (string – loajální / neloajální)
- Age – věk (int)
- Type of Travel – záměr cesty (string – byznys /osobní)
- Class – cestovní třída (string – Business / Eco / Eco Plus)
- Flight Distance – délka letu (int)
- Atributy míry uspokojenosti (int – [0-5])
    - Inflight wifi service – wifi připojení
    - Departure/Arrival time convenient – vhodná doba odletu / příletu
    - Ease of Online booking – kvalita online bookingu
    - Gate location – umístění brány
    - Food and drink – spokojenost s jídlem a pitím
    - Online boarding – online boarding (zařízení palubní vstupenky online)
    - Seat comfort – pohodlí sedadel
    - Inflight entertainme – zábava během letu
    - On-board service – obsluha při nalodění
    - Leg room service – prostor pro nohy
    - Baggage handling – zacházení se zavazadly
    - Checkin service – odbavení zavazadel
    - Inflight service – obsluha za letu
    - Cleanliness – čistota letadla
- Departure Delay in Minutes – zpoždění odletu v minutách (int)
- Arrival Delay in Minutes – zpoždění příletu v minutách (int)
- Satisfaction – celková spokojenost s aerolinkou (string – satisfied / neutral or dissatisfied)

2 Dolovací úlohy

2.1 Asociace mezi jednotlivými službami a spokojeností

Popis: Úloha zaměřená na zjištění, které poskytované služby mají největší vliv na celkovou spokojenost cestujících. V rámci úlohy budeme hledat souvislost mezi kvalitami služeb a celkovou spokojeností.

Použití: Letecká společnost může využít výsledky této analýzy ke zlepšení nebo prioritizaci daných důležitých služeb, za účelem zvýšení průměrné spokojenosti cestujících.

Metody: Asociační pravidla, Rozhodovací stromy

2.2 Analýza třídy letu v závislosti na věku a pohlaví cestujícího

Popis: Tato úloha se zaměřuje na analýzu, jak věk a pohlaví cestujícího ovlivňuje volbu cestovní třídy. Cílem je zjistit, zda existují nějaké demografické vzory v preferencích třídy.

Použití: Letecká společnost může využít výsledky analýzy, k vytváření cílených marketingových kampaních.

Metody: Klasifikace

Řešení úloh

3 Charakteristiky datové sady

- V datové sadě je celkově 103436 záznamů.
- Rozdělení můžu a žen je blízké polovičnímu, kde muži lehce převyšují (50.75 %).
- Obsahuje redundantní číslování řádku, které je v následujících úlohách odstraněno. Jako identifikátor byl použit atribut "id"
- Třída celkové spokojenosti není vyvážená [56.66 %, 43.34 %].
- Přes 80 % pasažérů jsou věrní zákazníci.
- Skoro 70 % letů jsou služební cesty.

4 Asociace mezi jednotlivými službami a spokojeností

4.1 Popis

Základem této úlohy je prozkoumání korelace jednotlivých služeb s výslednou spokojeností za celý let. Dále také bude obsahovat klasifikační analýzu spokojenosti na základě služeb. Výsledné informace z těchto dvou průzkumů by se potom dalo využít následovně.

- Které služby mají největší dopad na spokojenost, a tudíž jsou nejlepšími kandidáty na zlepšení.
- Odhadnutí spokojenosti letu v letadle na základě jeho služeb (zprůměrované ohodnocení služeb ze záznamů) a následné označení takových letů jako "doporučených".

Služby jsou ohodnoceny hodnotami [1,2,3,4,5]. Celková spokojenost je binární údaj. Pasažér je buď spokojen, nebo neutrální až nespokojený. Vyhodnocením této úlohy tedy bude náhled, které služby jsou nejdůležitější.

4.2 Předpřipravení dat

Všechny data se nacházeli v jednom souboru .csv, který obsahoval hlavičky atributů. Tedy samotný import do Rapid Mineru byl jednoduchý. Obsahuje ale zbytečnou informaci o číslu řádku, které je v každé další úloze také odstraněno. Pro identifikování záznamu byl použit "id".

Dalším krokem je selekce pouze potřebných atributů, což jsou jednotlivé služby, ID záznamu a údaj o celkové spokojenosti. Po selekci atributů neexistuje ani jeden záznam, které by měl chybějící hodnoty, tudíž není potřeba žádné doplňovat. A tím že jsou hodnoty v rozmezí 1-5, není zapotřebí ani normalizovat. Pouze bylo převedeno celkové uspokojení z:

- satisfied => 1
- neutral or dissatisfied => 0

4.3 Řešení

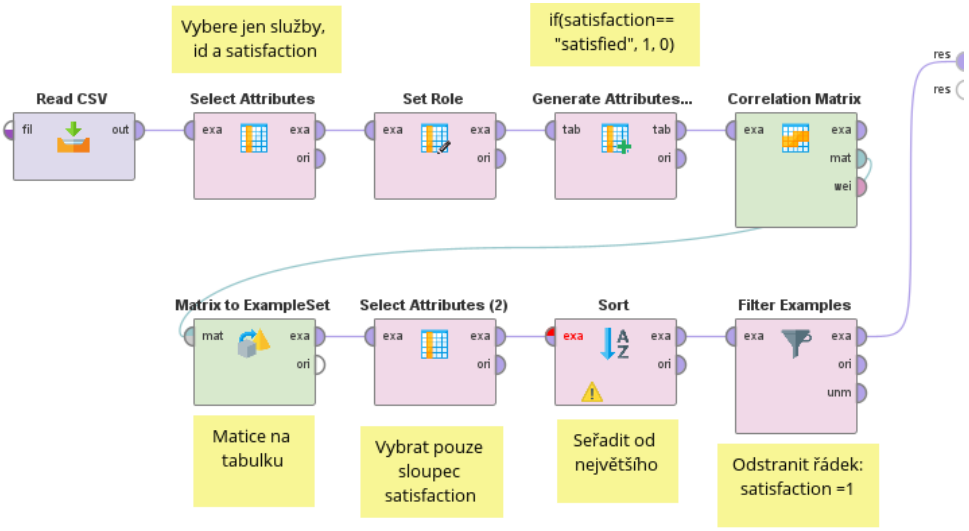
4.3.1 Ovlivnění kvality služeb na spokojenost

Pro výpočet korelace služeb a spokojenosti byla využita korelační matice[1]. Důležitý je sloupec s hodnotou spokojenost. Po filtraci a seřazení vypadá následovně:

Attributes	satisfaction
Online boarding	0.504
Inflight entertainment	0.398
Seat comfort	0.349
On-board service	0.322
Leg room service	0.313
Cleanliness	0.305
Inflight wifi service	0.284
Baggage handling	0.248
Inflight service	0.245
Checkin service	0.236
Food and drink	0.210
Ease of Online booking	0.172
Gate location	0.001
Departure/Arrival time convenient	-0.052

Tabulka: Korelace sužeb s atributem satisfaction

Tuto tabulku jsem získal následujícím procesem v Altair AI:

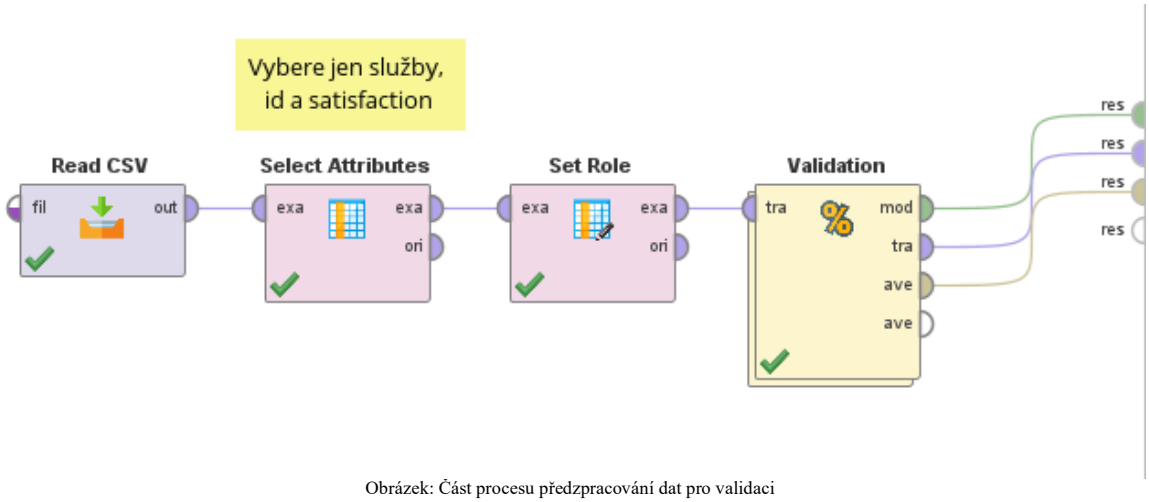


Obrázek: Proces korelace v Altair AI

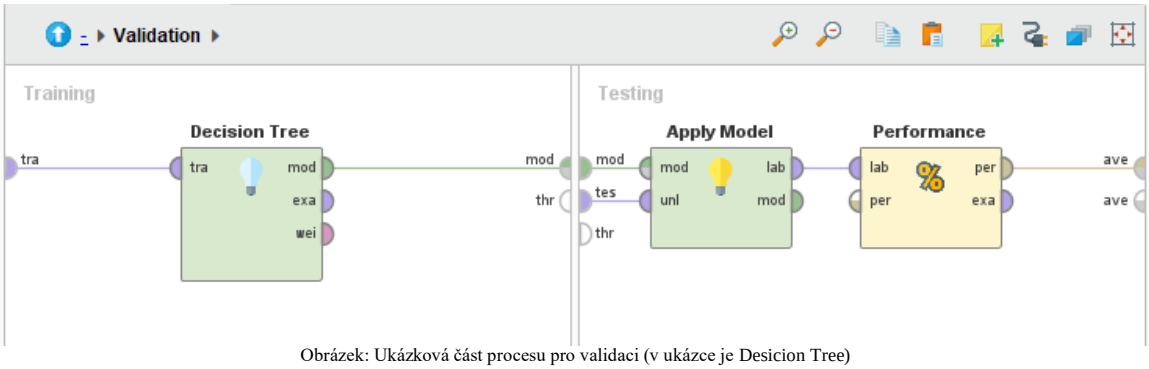
[1] Nebyl využit Weight by Correlation, protože špatně počítal hodnotu Departure/Arrival time convenient, která je záporná

4.3.2 Odhadnutí spokojenosti

Pro odhadování spokojenosti bylo vyzkoušeno několik různých metod, a to Desicion tree, Naive bayes, Deep Learning, Random Forest. Všechny ale měli stejnou zahajovací část (podobnou jako v kapitole *Ovlivnění kvality služeb na spokojenost*), a to načtení souboru, selekce služeb, id a spokojenosti, a nastavení role. Následně se data přeposlali do bloku **Validation,** který natrénuje daný model při rozdělení sady na trénovací / testovací s poměrem 0.75 (jednotlivé záznamy jsou do nich přidělovány náhodně) a nakonec ho aplikuje a vyhodnotí.



Obrázek: Část procesu předzpracování dat pro validaci



Obrázek: Ukázková část procesu pro validaci (v ukázce je Desicion Tree)

Je také důležité zdůraznit, že třídy spokojenosti nejsou vyvážené. Modely tedy budou mít sklon ke klasifikaci na hodnotu "neutral or dissatisfied"

Satisfaction	Count	Percentage
neutral or dissatisfied	58697	56.66%
satisfied	44897	43.34%

Tabulka: Rozložení atributu "Satisfaction"

4.3.2.1 Desicion tree

Jako první algoritmus na realizaci dané úlohy byl zvolen desicion tree. Po experimentálním ladění s následnou konfigurací dosáhl přesnosti **92.86 %**.

Konfigurace:

- criterion – gini index

- maximální hloubka – 12
- pruning – confidence = 0.15
- prepruning – vypnuto
- ostatní možnosti jsou ponechány v základním nastavení

Tabulka přesnosti pro Desicion Tree[2]:

	true neutral or dissatisfied	true satisfied	class precision
pred. neutral or dissatisfied	14089	1252	91.84%
pred. satisfied	602	10033	94.34%
class recall	95.90%	88.91%	

Tabulka: Přesnost modelu: Desicion Tree

4.3.2.2 Naive bayes

Přesnost řešení byla **78.30 %.**

	true neutral or dissatisfied	true satisfied	class precision
pred. neutral or dissatisfied	11341	2287	83.22%
pred. satisfied	3350	8998	72.87%
class recall	77.20%	79.73%	

Tabulka: Přesnost modelu: Naive bayes

4.3.2.3 Deep Learning

Jako reprezentativní model neuronové sítě byl zvolen operátor Deep Learning. Po ladění s následnou konfigurací dosáhl přesnosti **94.43 %**.

Konfigurace:

- aktivace – RectifierWithDropout[3]
- skryté vrstvy – celkem dvě vrstvy [100, 70]
- nastavení hodnoty dropout pro vrstvy - [0.1, 0.06]
- počet epoch - 20
- loss funkce – Cross entropy
- počet záznamů pro iteraci – automatický
- ostatní jsou ponechány v základním nastavení

	true neutral or dissatisfied	true satisfied	class precision
pred. neutral or dissatisfied	14249	1017	93.34%
pred. satisfied	442	10268	95.87%
class recall	96.99%	90.99%	

Tabulka: Přesnost modelu: Deep Learning

[2] class recall – měří, jak dobře model identifikuje pozitivní příklady pro každou třídu
class precision – měří, jaký podíl příkladů klasifikovaných jako určitá třída je správných

[3] Rectifier Linear Unit: vybere maximum z (0, x) kde x je vstupní hodnota

4.3.2.4 Random Forest

Přesnost řečení byla **94.16 %.**

Konfigurace:

- criterion – gini index
- počet stromů – 100
- maximální hloubka – 12
- pruning – confidence = 0.1
- prepruning – vypnuto
- ostatní jsou ponechány v základním nastavení

	true neutral or dissatisfied	true satisfied	class precision
pred. neutral or dissatisfied	14213	1039	93.19%
pred. satisfied	478	10246	95.54%
class recall	96.75%	90.79%	

Tabulka: Přesnost modelu: Desicion Tree

4.4 Závěr

4.4.1 Ovlivnění kvality služeb na spokojenost

Ze seřazené tabulky korelace jednotlivých služeb se spokojeností je zřejmé, které mají jak velký vliv, přičemž značně nejvyšší je Online boarding s hodnotou 0.504. Je potřeba ale vzít v potaz, že nevíme, jak pasažéři vyplňovali své hodnocení daných letů. Pokud bylo hodnocení provedeno online dotazníkem po vybídnutí emailem nebo aplikací, je pravděpodobné, že na něj spíše odpovídali technicky schopnější pasažéři, pro které je pak služba online boardingu důležitější.

4.4.2 Odhadnutí spokojenosti na základě služeb

Odhadnutí spokojenosti na základě služeb bylo prozkoumáno algoritmy Desicion tree, Naive bayes, Deep Learning a Random Forest, přičemž jako nejslibnější řešení se zde projevuje Deep Learning (94.43 %) a Random Forest (94.16 %). Dokonce i malý Desicion tree měl vysokou míru přesnosti (92.86 %). Díky rozdělení sady na trénovací / testovací s poměrem 0.75 lze očekávat, že se modely budou mít dobrou generalizaci.

Vliv na výsledky bude mít fakt, že zastoupení tříd v atributu Satisfaction není rovnoměrné. Každopádně vybrané modely klasifikují dobře. Tyto modely by se tedy dali využít v zadané úloze, která se týkala klasifikace letu / letadla na základě ohodnocení jeho služeb.

5 Analýza třídy letu v závislosti na věku a pohlaví cestujícího

5.1 Popis

Cílem úlohy je analyzovat vliv dostupných demografických atributů cestujícího, pro nás tedy věk a pohlaví, na výběr cestovní třídy (Business, Economy, Eco Plus). Tato analýza může odhalit demografické vzory, které letecké společnosti pomohou lépe cílit marketingové kampaně, jako například na kterou skupinu zákazníků by měli zaměřit reklamu, nebo jakým stylem by reklama zákazníky oslovit.

Jak bylo zmíněno výše, tak datová sada obsahuje 3 typy cestovních tříd:

- Eco Plus
- Eco
- Business

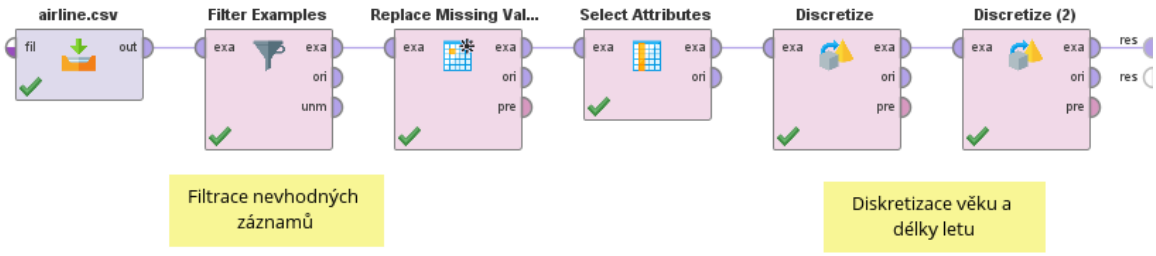
Tedy vhodnou dolovací úlohou je klasifikační analýza, kdy budeme na základě demografických údajů odhadovat, jakou třídu cestující zvolí. Vstupem této úlohy jsou dva parametry:

- Věk
- Pohlaví

Následně je úloha ještě doplněna o další dvě vlastnosti cestujícího a to:

- Účel cesty
- Délka letu

5.2 Předpřipravení dat



Obrázek: Schéma předzpracování dat v Altair AI

5.2.1 Filtrace dat a nahrazení chybějících numerických hodnot

Jako první je provedena filtrace datové sady, jsou provedeny dva druhy filtrace:

- Vyřazení záznamů s chybějícími kategorickými hodnotami
- Vyřazení záznamů s negativními numerickými hodnotami

Pro atributy “Class” a “Type of travel” byly vyfiltrovány všechny záznamy, které neobsahovaly tuto hodnotu. Podobně pro atributy “Age” a “Flight Distance” byly vyfiltrovány záznamy jejichž hodnota byla menší jak nula. Filtrace dat byla provedena pomocí operátoru “Filter Examples”.

Chybějící numerické hodnoty jsou nahrazeny průměrem datové sady za pomocí operátoru “Replace Missing Values”.

5.2.2 Výběr atributů

Následně je provedena selekce atributů, se kterými chceme pracovat. Pro naši analýzu nás zajímají atributy:

- Class
- Type of travel
- Age
- Gender
- Flight distance

5.2.3 Diskretizace věku

Dále se provádí diskretizace věku. Věk byl rozdělen do intervalů, kde jednotlivý interval pokrývá rozmezí 10 let. Výjimkou je poslední interval, který pokrývá cestující ve věkové skupině od 80 do 129 let. Diskretizace se provedla pomocí operátoru "Discretize by User Specification"

Intervaly:

- <0;9>
- <10;19>
- <20;29>
- ...
- <80;129>

5.2.4 Diskretizace délky letu

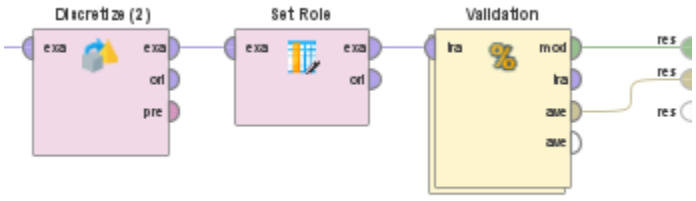
Délka letu je v datové sadě uvedena v kilometrech a byla rozdělena na tři skupiny

- Short - <0;1000>
- Medium - <1001;3000>
- Long – <3001; 3001+>

5.3 Řešení

Úloha byla nejdříve řešena s použitím dvou základních atributů a následně provedena i pro další dva rozšiřující atributy popsané v části 5.1. Pro řešení úlohy byly použity dva druhy prediktivních modelů, a to Bayesovská a stromová klasifikace. Následně byla provedena analýza rozšířená o další dva atributy, opět pomocí Bayesovského modelu.

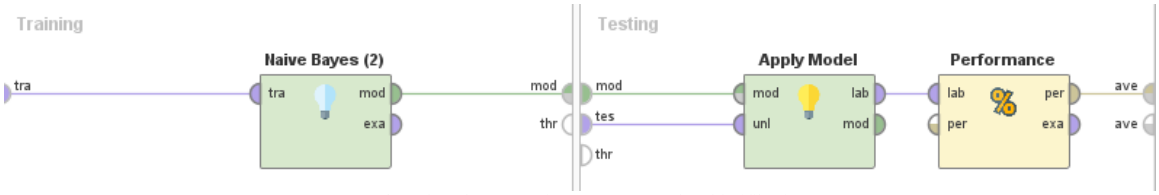
Samotné vykonání modelů bylo provedeno pomocí operátoru "Validation", u kterého byl nastaven poměr dat na 0.75. Tedy 75 % dat bylo použito jako trénovací data a zbylých 25 % jako testovací. Před posláním dat do operátoru "Validation" bylo ještě zapotřebí nastavit cílový atribut, to bylo zajištěno pomocí operátoru "Set Role".



Obrázek: Schéma zapojení Validation v Altair AI

5.3.1 Bayesovská klasifikace

Tato klasifikace byla provedena za pomocí operátoru “Naive Bayes”, který za vstup bere sadu trénovacích dat. Výstup tohoto modelu byl poslán na operátor “Apply Model” společně s testovacími daty. Statistická výkonost modelu byla pak zpracována operátorem “Performance”, jehož výstupem bylo ohodnocení přesnosti modelu, jeho hlavní kritérium bylo tedy nastaveno na “accuracy”. Tato klasifikace také vygenerovala několik užitečných grafů rozložení hustoty na základě vstupních atributů, které jsou popsané níže.



Obrázek: Schéma zapojení pro Bayesovskou klasifikaci

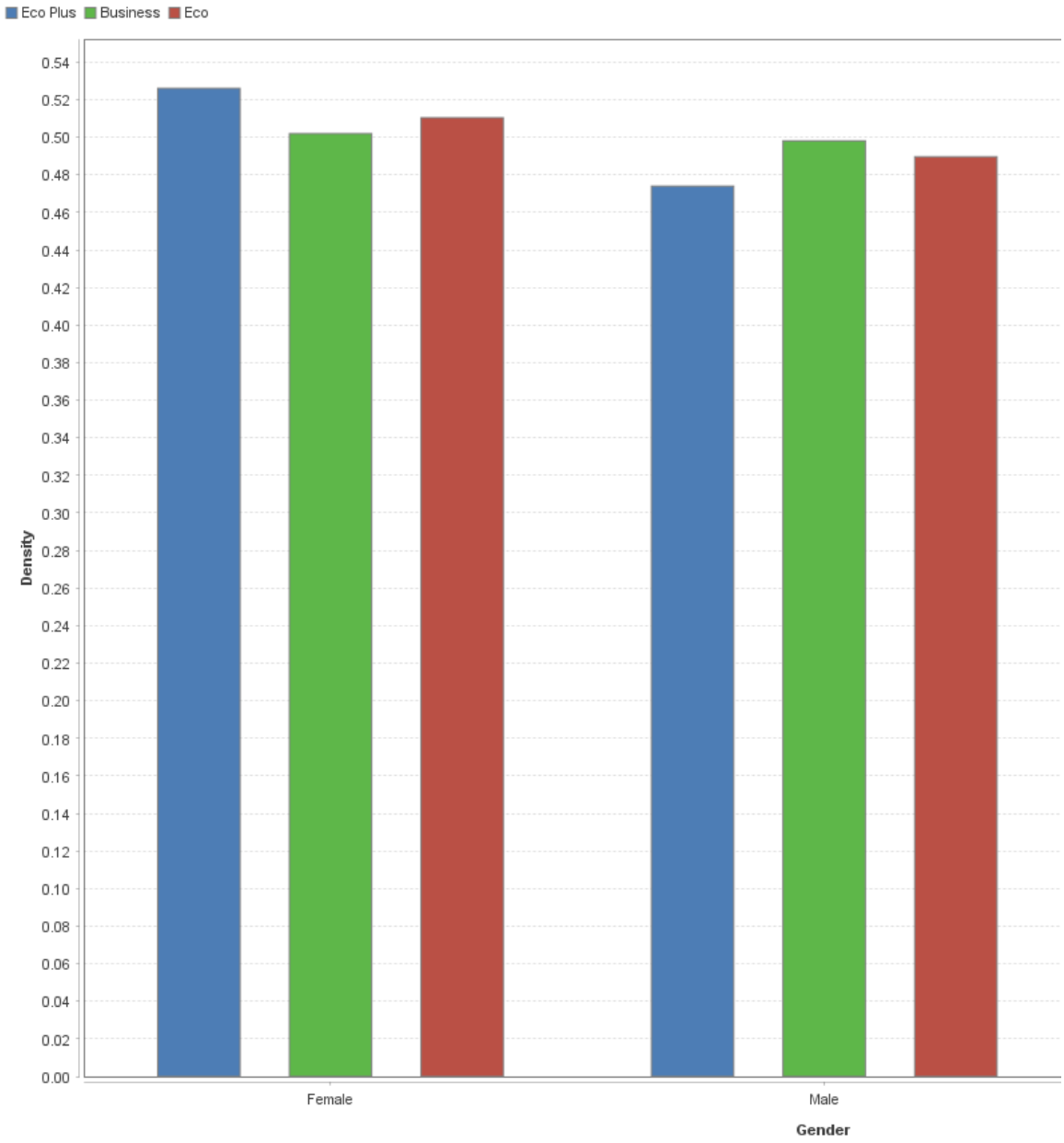
Odhad cestovní třídy pomocí tohoto modelu měl následující výsledky dosáhl průměrné přesnosti 58,13 %. V tabulce popisující přesnost odhadu si můžeme povšimnout, že třída Eco Plus má přesnost 0 %, to je pravděpodobně způsobeno velice malým výskytem cestujících danou třídou (7494 z 103904).

	true Eco Plus	true Business	true Eco	class precision
pred. Eco Plus	0	0	0	0,00%
pred. Business	1044	9075	5661	57,51%
pred. Eco	829	3341	6025	59,10%
class recall	0,00%	73,09%	51,56%	

Tabulka: Přesnost Bayesovské klasifikace

5.3.1.1 Graf rozložení hustoty cestovních tříd podle pohlaví

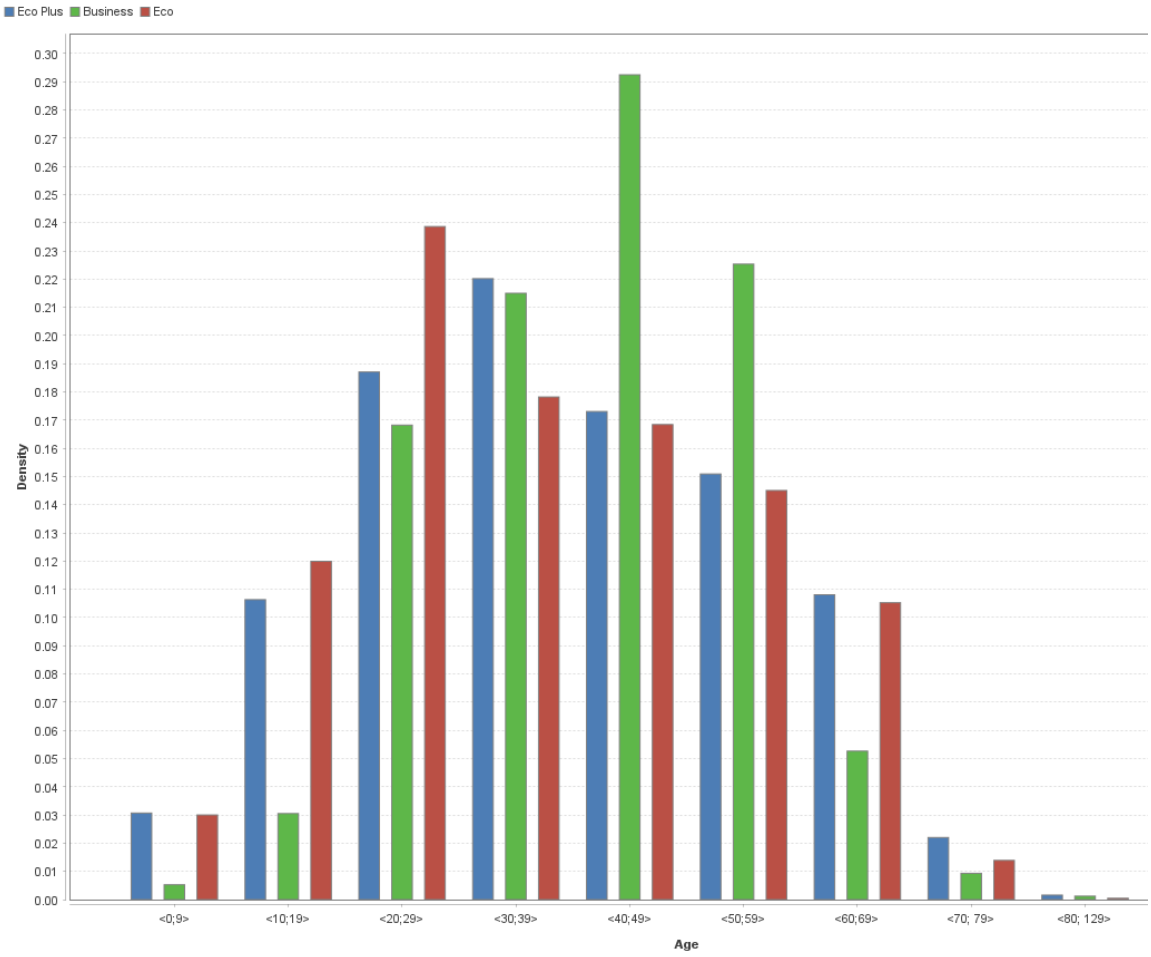
Z tohoto grafu lze vypozorovat, že rozložení tříd mezi pohlavími je skoro vyrovnané a s velkou pravděpodobností tak nehraje roli při rozhodování.



Obrázek: Graf rozložení hustoty cestovních tříd podle pohlaví cestujícího

5.3.1.2 Graf rozložení hustoty cestovních tříd podle věku

Z tohoto grafu lze pozorovat, že největší část letů v Business třídě tvoří cestující ve věku 40-59 let, konrétně tedy 51,7 %. Mladší cestující převážně preferují Eco třídu. Také lze z tohoto grafu pozorovat, že většinu celkových letů tvoří cestující ve věku 20-59 let.

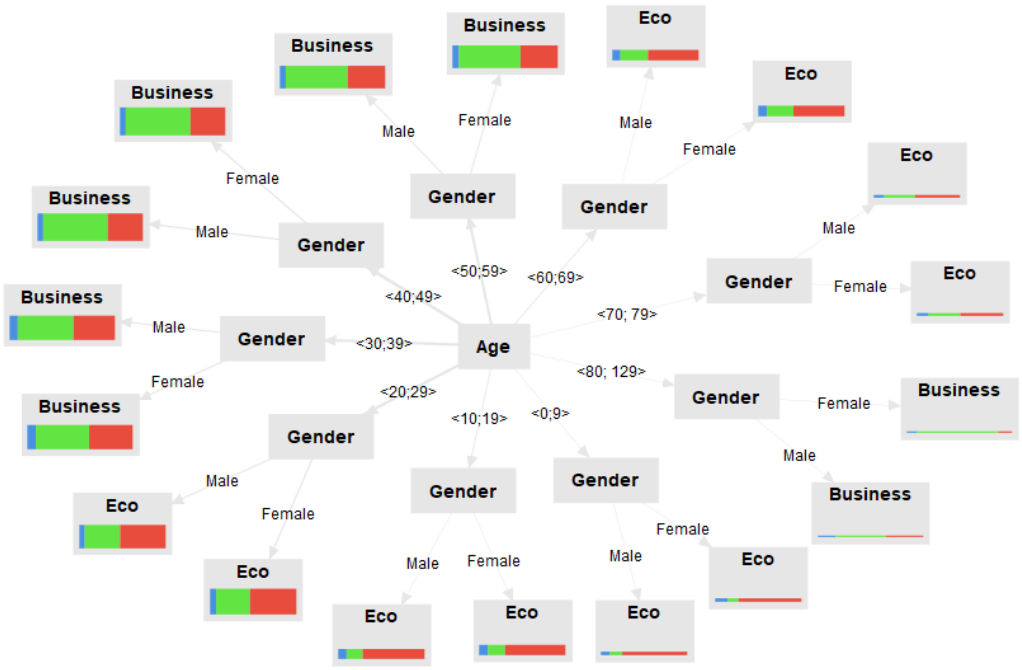


Obrázek: Graf rozložení hustoty cestovních tříd podle věku cestujícího

5.3.2 Tree klasifikace

Postup pro provedení stromové klasifikace byl velice podobný jako u Bayesovské. Jako model byl zvolen operátor “Decision Tree”.

Tato klasifikace dosáhla stejné přesnosti jako Bayesovská, tedy 58,13 %.



Obrázek: Rozhodovací strom zobrazující volbu cestovní třídy na základě věku a pohlaví

Ze samotného rozhodovacího stromu můžeme opět vypozorovat, že pohlaví cestujícího při výběru třídy není nijak důležitým faktorem. Naopak co se věkové kategorie týče, tak zde můžeme pozorovat různé preference tříd mezi skupinami. Tato analýza se tedy shoduje s výsledky Bayesovské klasifikace.

5.3.3 Bayesovská klasifikace – rozšířená verze

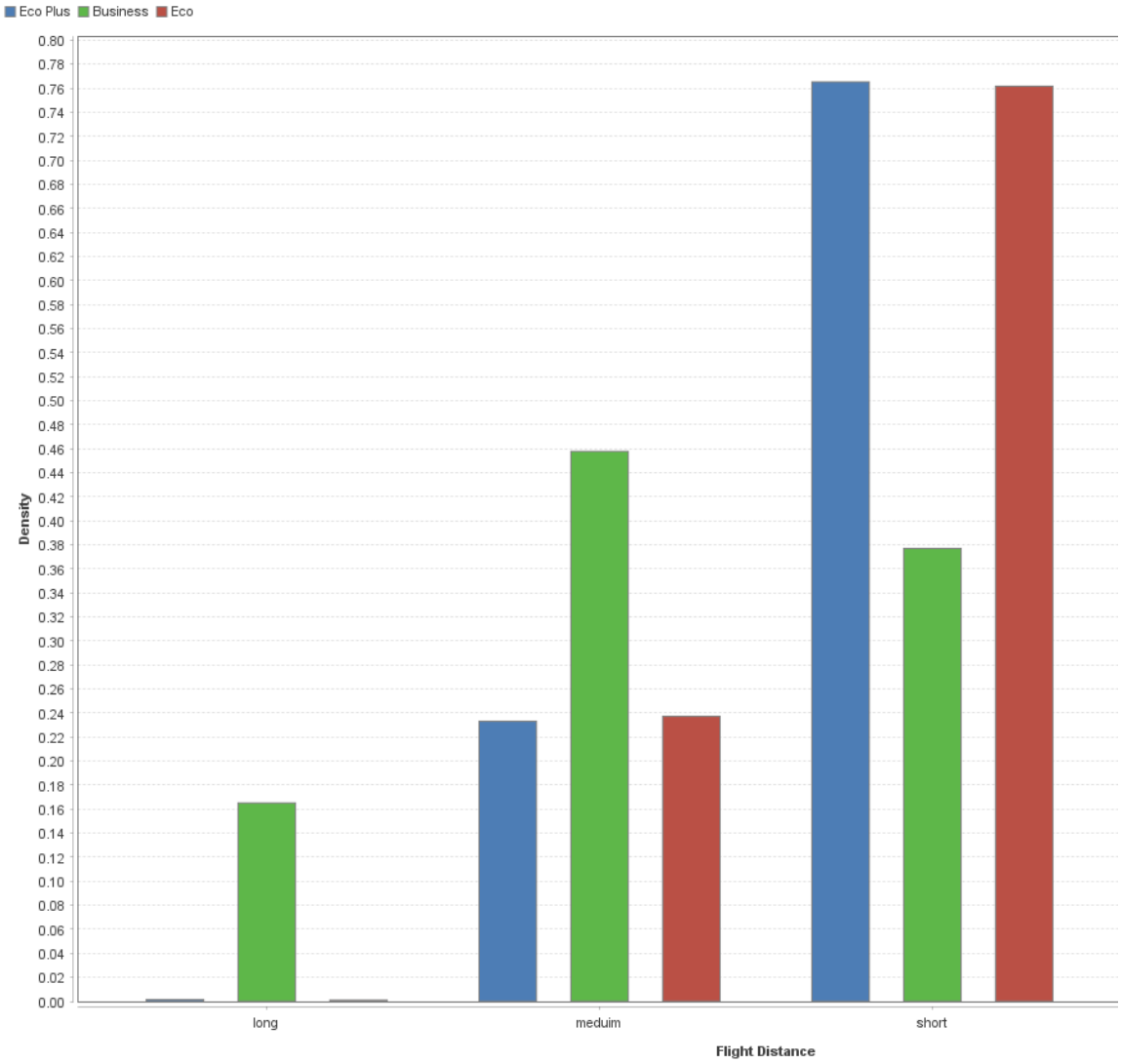
Tato klasifikace byla rozšířena od dodatečné atributy, čímž se podařilo dosáhnout vyšší přesnosti odhadu cestovní třídy. Průměrná přesnost této klasifikace byla 73,43 %. Z toho můžeme vydedukovat, že čím více znalostí o cestujícím známé, tím lepší odhad, jakou cestovní třídu si zvolí můžeme provést. Bohužel datová sada neobsahuje žádné dodatečné informace o cestujícím, které bychom v rámci této analýzy mohli využít.

	true Eco Plus	true Business	true Eco	class precision
pred. Eco Plus	0	0	0	0,00%
pred. Business	668	10591	3203	73,23%
pred. Eco	1205	1825	8483	73,68%
class recall	0,00%	85,30%	72,59%	

Tabulka: Přesnost Bayesovské klasifikace

5.3.3.1 Graf rozložení hustoty cestovních tříd podle délky letu

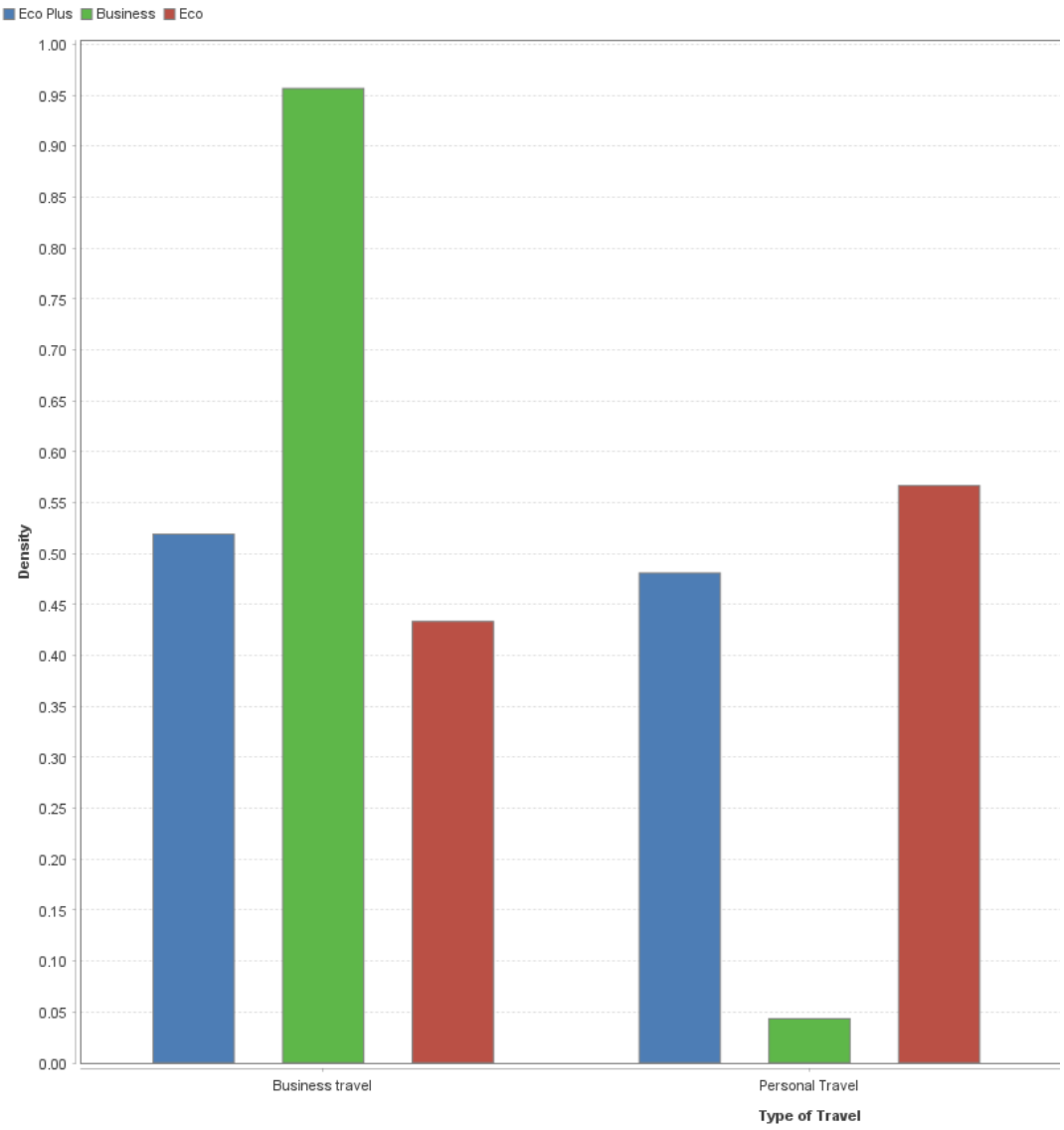
Z tohoto grafu lze vypozorovat, že většina cestujících volí Eco třídu při kratších letech, což je očekávaný jev. Naopak při delších letech cestující upřednostňují Business třídu, pravděpodobně kvůli lepšímu komfortu, který může dlouhý let zpříjemnit. Dalo by se tedy usoudit, že pokud cestující bude podnikat dlouhou trasu, pak bude více preferovat komfort než cenu.



Obrázek: Graf rozložení hustoty cestovních tříd podle délky letu

5.3.3.2 Graf rozložení hustoty cestovních tříd podle typu cesty

Z tohoto grafu lze pozorovat, že lety s za pracovním účelem tvoří většinu letů v Business třídě, pravděpodobně, protože tato třída poskytuje lepší komfort na práci během letu.



Obrázek: Graf rozložení hustoty cestovních tříd podle účelu cesty

5.3.4 Bayesovská klasifikace – oversampled

Aby byl vyřešen problém s nulovou predikcí třídy Eco Plus, tak byl do Altair AI stažen dodatečný balíček, který poskytuje operátor "SMOTE". Tento operátor automaticky vytvoří dodatečné záznamy pro nedostatečně zastoupené cestovní třídy, v našem případě Eco Plus, aby rozložení mezi jednotlivými třídami bylo rovnoměrné. Vytvoření nových záznamů se dělá pomocí duplikování existujících, tudíž výše uvedené grafy hustoty nebudou nijak ovlivněny.

Klasifikace se zlepšila v ohledu predikce Eco Plus, ta již nenabývá nulových hodnot, ale dosahuje přesnosti odhadu 44,56 %. Celkově se ovšem zhoršila průměrná přesnost predikce cestovní třídy na 54,07 %, a to i přestože jsou do analýzy zahrnuty dodatečné atributy z předchozí klasifikace. Což opět potvrzuje, že pro takovou klasifikaci by bylo zapotřebí znát více informací o cestujícím.

	true Eco Plus	true Business	true Eco	class precision
pred. Eco Plus	5133	2202	4184	44,56%
pred. Business	3094	9376	2264	63,64%
pred. Eco	4189	838	5238	51,03%
class recall	41,34%	75,52%	44,82%	

Tabulka: Přesnost Bayesovské klasifikace

5.4 Závěr

Celkově lze z analýzy vyhodnotit, že dva atributy popisující cestujícího nejsou dostatečné a ani po přidání dvou dodatečných atributů nebyla přesnost příliš přesvědčivá. Pokud bychom tedy chtěli cestující spolehlivě klasifikovat do cestovních tříd bylo by zapotřebí o nich znát více, jako například jich aktuální pracovní stav (nezaměstnaný/zaměstnaný/student) nebo jestli třeba cestoval sám nebo ve skupině.

Pokud bychom se tedy ale drželi výsledků naší analýzy, převážně tedy výsledků Bayesovské klasifikace, tak lze usoudit preference vyšší cestovní třídy pro cestující ve věku 40-59 let, jejichž účel letu je spíše pracovní. Pro marketingovou kampaň by to mohlo například znamenat, že pokud chceme propagovat Business třídu, tak by reklama měla vyobrazovat zaměstnance, ve středním věku, kteří jsou na pracovním letu, cestují Business třídou a mají tak klid na svou práci. Stejným způsobem lze postupovat pro určení zaměření reklam na třídu Eco nebo Eco Plus.