CS 584: Machine Learning

Pradyot Mayank(A20405826)
Spring 2019 Assignment 3

Question 1 (50 points)

You will use the CART algorithm to build profiles of credit card holders. The data is the CustomerSurveyData.csv. The analysis specifications are:

Target Variable

- **CarOwnership**. The type of car ownership. This variable has three non-missing categories which are *Leased*, *None*, and *Own*.
- Drop all missing values in the target variable.

Nominal Predictors

- **CreditCard**. The type of credit card held. This variable has five categories which are *American Express*, *Discover*, *MasterCard*, *Others*, and *Visa*.
- **JobCategory**. The category of the job held. This variable has six non-missing categories which are *Agriculture*, *Crafts*, *Labor*, *Professional*, *Sales*, and *Service*.
- Recode all the missing values into the *Missing* category.

You will use the Entropy metric as the splitting criterion. You may want to write a Python program to assist you in answering the questions.

a) (5 points). What is the Entropy metric for the root node?

Root node Entropy=1.0744900777690842

b) (5 points). How many possible binary-splits that you can generate from the CreditCard predictor?

K=5
No of binary splits=2^{k-1}-1=15

c) (10 points). Calculate the Entropy metric for each possibly binary split that you can generate from the CreditCard predictor. List your answers in a table. The table should have three columns: an index of the split, the contents of the two branches, and the split entropy metric.

SplitIndex	Left_Child	Right_Child	SplitEntropy	Counter
[1.0468094317797334, 1.1479670231237966]	['Visa']	['American Express', 'Discover', 'MasterCard', 'Others']	1.072038135	1
[1.0725070489912254, 1.075008859357403]	['American Express']	['Discover', 'MasterCard', 'Others', 'Visa']	1.073000406	1
[1.1085058090287196, 0.973200661032307]	['Discover']	['American Express', 'MasterCard', 'Others', 'Visa']	1.072135785	1
[1.0689643982639212, 1.174264078521314]	['Others']	['American Express', 'Discover', 'MasterCard', 'Visa']	1.073660764	1
[1.0760277584181488, 1.069359232936541]	['MasterCard']	['American Express', 'Discover', 'Others', 'Visa']	1.074427312	1
[1.0348662734428489, 1.1211075887224382]	['American Express', 'Visa']	['Discover', 'MasterCard', 'Others']	1.073381645	2
[1.0852881500406055, 1.0638912768034012]	['Discover', 'Visa']	['American Express', 'MasterCard', 'Others']	1.07420029	2
[1.0369375321912813, 1.1522460232150684]	['Others', 'Visa']	['American Express', 'Discover', 'MasterCard']	1.070838229	2
[1.0351668498221591, 1.1109476274999246]	['MasterCard', 'Visa']	['American Express', 'Discover', 'Others']	1.072253962	2
[1.1167776104815863, 1.018286062470057]	['American Express', 'Discover']	['MasterCard', 'Others', 'Visa']	1.070880549	2
[1.065626118599882, 1.0992153713223232]	['American Express', 'Others']	['Discover', 'MasterCard', 'Visa']	1.073748	2
[1.0738266333163597, 1.0739768609820293]	['American Express', 'MasterCard']	['Discover', 'Others', 'Visa']	1.073892313	2
[1.1034337362362638, 1.007246961317521]	['Discover', 'Others']	['American Express', 'MasterCard', 'Visa']	1.073288801	2
[1.1269534892368755, 1.0203280170186693]	['Discover', 'MasterCard']	['American Express', 'Others', 'Visa']	1.072702449	2
[1.0686782222097913, 1.087853923191551]	['MasterCard', 'Others']	['American Express', 'Discover', 'Visa']	1.074135627	2
[1.087853923191551, 1.0686782222097913]	['American Express', 'Discover', 'Visa']	['MasterCard', 'Others']	1.074135627	3
[1.0203280170186693, 1.1269534892368755]	['American Express', 'Others', 'Visa']	['Discover', 'MasterCard']	1.072702449	3
[1.007246961317521, 1.1034337362362638]	['American Express', 'MasterCard', 'Visa']	['Discover', 'Others']	1.073288801	3
[1.0739768609820293, 1.0738266333163597]	['Discover', 'Others', 'Visa']	['American Express', 'MasterCard']	1.073892313	3
[1.0992153713223232, 1.065626118599882]	['Discover', 'MasterCard', 'Visa']	['American Express', 'Others']	1.073748	3
[1.018286062470057, 1.1167776104815863]	['MasterCard', 'Others', 'Visa']	['American Express', 'Discover']	1.070880549	3
[1.1109476274999246, 1.0351668498221591]	['American Express', 'Discover', 'Others']	['MasterCard', 'Visa']	1.072253962	3
[1.1522460232150684, 1.0369375321912813]	['American Express', 'Discover', 'MasterCard']	['Others', 'Visa']	1.070838229	3
[1.0638912768034012, 1.0852881500406055]	['American Express', 'MasterCard', 'Others']	['Discover', 'Visa']	1.07420029	3
[1.1211075887224382, 1.0348662734428489]	['Discover', 'MasterCard', 'Others']	['American Express', 'Visa']	1.073381645	3
[1.069359232936541, 1.0760277584181488]	['American Express', 'Discover', 'Others', 'Visa']	['MasterCard']	1.074427312	4
[1.174264078521314, 1.0689643982639212]	['American Express', 'Discover', 'MasterCard', 'Visa']	['Others']	1.073660764	4
[0.973200661032307, 1.1085058090287196]	['American Express', 'MasterCard', 'Others', 'Visa']	['Discover']	1.072135785	4
[1.075008859357403, 1.0725070489912254]	['Discover', 'MasterCard', 'Others', 'Visa']	['American Express']	1.073000406	4
[1.1479670231237966, 1.0468094317797334]	['American Express', 'Discover', 'MasterCard', 'Others']	['Visa']	1.072038135	4

d) (5 points). What is the optimal split for the CreditCard predictor?

Optimal Split entropy for card: 1.0708382285522746

['Others', 'Visa'] and ['American Express', 'Discover', 'MasterCard']

e) (5 points). How many possible binary-splits that you can generate from the JobCategory predictor?

K=7

No of binary splits=2^{k-1}-1=63

f) (10 points). Calculate the Entropy metric for each possibly binary split that you can generate from the JobCategory predictor. List your answers in a table. The table should have three columns: an index of the split, the contents of the two branches, and the split Entropy metric.

SplitIndex	Left_Child	Right_Child	SplitEntropy	Counter
[1.0833972877916547, 1.0460584911356179]	['Professional']	['Agriculture', 'Crafts', 'Labor', 'Missing', 'Sales', 'Service']	1.07309178	1
[1.0759755834161266, 0.35335933502142136]	['Missing']	['Agriculture', 'Crafts', 'Labor', 'Professional', 'Sales', 'Service']	1.073807735	1
[1.064580524786143, 1.1656875625039393]	['Crafts']	['Agriculture', 'Labor', 'Missing', 'Professional', 'Sales', 'Service']	1.073720601	1
[1.0723009830941819, 1.1194267388242052]	['Agriculture']	['Crafts', 'Labor', 'Missing', 'Professional', 'Sales', 'Service']	1.074299115	1
[1.0565487841441326, 1.105165681538302]	['Sales']	['Agriculture', 'Crafts', 'Labor', 'Missing', 'Professional', 'Service']	1.07244651	1
[1.083648134315872, 1.0141965601531442]	['Labor']	['Agriculture', 'Crafts', 'Missing', 'Professional', 'Sales', 'Service']	1.074119378	1
[1.08146199500917, 1.0223048390077045]	['Service']	['Agriculture', 'Crafts', 'Labor', 'Missing', 'Professional', 'Sales']	1.074126508	1
[1.0854689598987637, 1.040929686432672]	['Missing', 'Professional']	['Agriculture', 'Crafts', 'Labor', 'Sales', 'Service']	1.073042503	2
[1.0701243089471246, 1.0772500145098465]	['Crafts', 'Professional']	['Agriculture', 'Labor', 'Missing', 'Sales', 'Service']	1.072735167	2
[1.0810283932597915, 1.057696034205267]	['Agriculture', 'Professional']	['Crafts', 'Labor', 'Missing', 'Sales', 'Service']	1.07359937	2
[1.0628364325882558, 1.0819741455142986]	['Professional', 'Sales']	['Agriculture', 'Crafts', 'Labor', 'Missing', 'Service']	1.074376473	2
[1.098775860703346, 1.0368434189492128]	['Labor', 'Professional']	['Agriculture', 'Crafts', 'Missing', 'Sales', 'Service']	1.073185376	2
[1.0947928997116794, 1.0388985423848909]	['Professional', 'Service']	['Agriculture', 'Crafts', 'Labor', 'Missing', 'Sales']	1.072435157	2
[1.0662013859582715, 1.1487127055891695]	['Crafts', 'Missing']	['Agriculture', 'Labor', 'Professional', 'Sales', 'Service']	1.073907943	2
[1.0738491304787368, 1.0848976516491609]	['Agriculture', 'Missing']	['Crafts', 'Labor', 'Professional', 'Sales', 'Service']	1.074350733	2
[1.0587133975770677, 1.1005520518185585]	['Missing', 'Sales']	['Agriculture', 'Crafts', 'Labor', 'Professional', 'Service']	1.072520153	2
[1.085382764827361, 1.0041174226279344]	['Labor', 'Missing']	['Agriculture', 'Crafts', 'Professional', 'Sales', 'Service']	1.073989364	2
[1.0831687384502102, 1.011181396991377]	['Missing', 'Service']	['Agriculture', 'Crafts', 'Labor', 'Professional', 'Sales']	1.074026346	2
[1.0616717625302443, 1.1533047155511529]	['Agriculture', 'Crafts']	['Labor', 'Missing', 'Professional', 'Sales', 'Service']	1.073840619	2
[1.0381822330626997, 1.1205464091449882]	['Crafts', 'Sales']	['Agriculture', 'Labor', 'Missing', 'Professional', 'Service']	1.07256104	2
[1.0732011608648113, 1.0787178435864888]	['Crafts', 'Labor']	['Agriculture', 'Missing', 'Professional', 'Sales', 'Service']	1.074456758	2
[1.0708182452602073, 1.0858785470267547]	['Crafts', 'Service']	['Agriculture', 'Labor', 'Missing', 'Professional', 'Sales']	1.074047174	2
[1.0516215330913017, 1.106818019118475]	['Agriculture', 'Sales']	['Crafts', 'Labor', 'Missing', 'Professional', 'Service']	1.072011115	2
[1.0815760236756207, 1.0405268873227713]	['Agriculture', 'Labor']	['Crafts', 'Missing', 'Professional', 'Sales', 'Service']	1.074203599	2
[1.079346323707148, 1.0496159707696378]	['Agriculture', 'Service']	['Crafts', 'Labor', 'Missing', 'Professional', 'Sales']	1.074399193	2
[1.0664137382279368, 1.0796798485518313]	['Labor', 'Sales']	['Agriculture', 'Crafts', 'Missing', 'Professional', 'Service']	1.072571867	2
[1.0640981251431965, 1.084832738893362]	['Sales', 'Service']	['Agriculture', 'Crafts', 'Labor', 'Missing', 'Professional']	1.073449436	2
[1.0933546901350217, 1.0186772027345765]	['Labor', 'Service']	['Agriculture', 'Crafts', 'Missing', 'Professional', 'Sales']	1.07384893	2
[1.0724682552792126, 1.0732636289741269]	['Crafts', 'Missing', 'Professional']	['Agriculture', 'Labor', 'Sales', 'Service']	1.072762066	3
[1.0832250496205886, 1.053175602362247]	['Agriculture', 'Missing', 'Professional']	['Crafts', 'Labor', 'Sales', 'Service']	1.073567157	3
[1.0665485706655906, 1.0795077429354574]	['Missing', 'Professional', 'Sales']	['Agriculture', 'Crafts', 'Labor', 'Service']	1.074401829	3
[1.1013635685496768, 1.0334150018958645]	['Labor', 'Missing', 'Professional']	['Agriculture', 'Crafts', 'Sales', 'Service']	1.073083375	3
[1.0973163471734606, 1.0353672141402663]	['Missing', 'Professional', 'Service']	['Agriculture', 'Crafts', 'Labor', 'Sales']	1.072350847	3
[1.0664289628782015, 1.0828457263946665]	['Agriculture', 'Crafts', 'Professional']	['Labor', 'Missing', 'Sales', 'Service']	1.073140136	3
[1.029287031473372, 1.093727830044641]	['Crafts', 'Professional', 'Sales']	['Agriculture', 'Labor', 'Missing', 'Service']	1.073970281	3
[1.0848685051635965, 1.061528458852803]	['Crafts', 'Labor', 'Professional']	['Agriculture', 'Missing', 'Sales', 'Service']	1.073114458	3
[1.0801497530150346, 1.0637767567882996]	['Crafts', 'Professional', 'Service']	['Agriculture', 'Labor', 'Missing', 'Sales']	1.072120436	3
[1.0552276269317562, 1.0846330429881488]	['Agriculture', 'Professional', 'Sales']	['Crafts', 'Labor', 'Missing', 'Service']	1.074205882	3
[1.097060279151978, 1.0455566089719206]	['Agriculture', 'Labor', 'Professional']	['Crafts', 'Missing', 'Sales', 'Service']	1.073595207	3
[1.0928477755340176, 1.0479383445566497]	['Agriculture', 'Professional', 'Service']	['Crafts', 'Labor', 'Missing', 'Sales']	1.072979843	3
[1.0869271931118374, 1.069827265560206]	['Labor', 'Professional', 'Sales']	['Agriculture', 'Crafts', 'Missing', 'Service']	1.074269827	3
[1.0803059719032344, 1.0722954900062025]	['Professional', 'Sales', 'Service']	['Agriculture', 'Crafts', 'Labor', 'Missing']	1.074482352	3
[1.1171463666766375, 1.0335086018380089]	['Labor', 'Professional', 'Service']	['Agriculture', 'Crafts', 'Missing', 'Sales']	1.072216159	3
[1.06336735070003, 1.141696498492564]	['Agriculture', 'Crafts', 'Missing']	['Labor', 'Professional', 'Sales', 'Service']	1.074004449	3
[1.0406431419295887, 1.1168839156768722]	['Crafts', 'Missing', 'Sales']	['Agriculture', 'Labor', 'Professional', 'Service']	1.072694763	3
[1.075124192156846, 1.072250544494806]	['Crafts', 'Labor', 'Missing']	['Agriculture', 'Professional', 'Sales', 'Service']	1.074461529	3
[1.0727065649901748, 1.0790146722738436]	['Crafts', 'Missing', 'Service']	['Agriculture', 'Labor', 'Professional', 'Sales']	1.074077948	3
[1.0539185960148778, 1.1027261973336537]	['Agriculture', 'Missing', 'Sales']	['Crafts', 'Labor', 'Professional', 'Service']	1.072094547	3
[1.0833968299697654, 1.0325916220970428]	['Agriculture', 'Labor', 'Missing']	['Crafts', 'Professional', 'Sales', 'Service']	1.074119799	3
[1.0811365804806505, 1.0410218302426482]	['Agriculture', 'Missing', 'Service']	['Crafts', 'Labor', 'Professional', 'Sales']	1.074341142	3
[1.0691513112283386, 1.0764816952177274]	['Labor', 'Missing', 'Sales']	['Agriculture', 'Crafts', 'Professional', 'Service']	1.072576067	3
[1.0667719532421294, 1.0815273656592308]	['Missing', 'Sales', 'Service']	['Agriculture', 'Crafts', 'Labor', 'Professional']	1.07347091	3
[1.0953977462439153, 1.01333671386026]	['Labor', 'Missing', 'Service']	['Agriculture', 'Crafts', 'Professional', 'Sales']	1.073717221	3
[1.0308769235107282, 1.1204722327452645]	['Agriculture', 'Crafts', 'Sales']	['Labor', 'Missing', 'Professional', 'Service']	1.072072847	3
[1.0702661113667726, 1.085804291396592]	['Agriculture', 'Crafts', 'Labor']	['Missing', 'Professional', 'Sales', 'Service']	1.07446142	3
[1.067836320199731, 1.0927177278350615]	['Agriculture', 'Crafts', 'Service']	['Labor', 'Missing', 'Professional', 'Sales']	1.074225866	3
[1.0447060737947207, 1.095559927326196]	['Crafts', 'Labor', 'Sales']	['Agriculture', 'Missing', 'Professional', 'Service']	1.072909621	3
[1.0424132388295584, 1.0998465913508486]	['Crafts', 'Sales', 'Service']	['Agriculture', 'Labor', 'Missing', 'Professional']	1.073507656	3
[1.08232263798382, 1.0593528409154878]	['Crafts', 'Labor', 'Service']	['Agriculture', 'Missing', 'Professional', 'Sales']	1.074246457	3
[1.060824547807569, 1.0830890061851488]	['Agriculture', 'Labor', 'Sales']	['Crafts', 'Missing', 'Professional', 'Service']	1.072103722	3
[1.0586487301995884, 1.0878981085533663]	['Agriculture', 'Sales', 'Service']	['Crafts', 'Labor', 'Missing', 'Professional']	1.073080373	3
[1.091563683148005, 1.0339639874712196]	['Agriculture', 'Labor', 'Service']	['Crafts', 'Missing', 'Professional', 'Sales']	1.074076416	3
[1.079242179444215, 1.0689339249527698]	['Labor', 'Sales', 'Service']	['Agriculture', 'Crafts', 'Missing', 'Professional']	1.073178864	3
[1.0689339249527698, 1.079242179444215]	['Agriculture', 'Crafts', 'Missing', 'Professional']	['Labor', 'Sales', 'Service']	1.073178864	4
[1.0339639874712196, 1.091563683148005]	['Crafts', 'Missing', 'Professional', 'Sales']	['Agriculture', 'Labor', 'Service']	1.074076416	4
[1.0878981085533663, 1.0586487301995884]	['Crafts', 'Labor', 'Missing', 'Professional']	['Agriculture', 'Sales', 'Service']	1.073080373	4
[1.0830890061851488, 1.060824547807569]	['Crafts', 'Missing', 'Professional', 'Service']	['Agriculture', 'Labor', 'Sales']	1.072103722	4
[1.0593528409154878, 1.08232263798382]	['Agriculture', 'Missing', 'Professional', 'Sales']	['Crafts', 'Labor', 'Service']	1.074246457	4
[1.0998465913508486, 1.0424132388295584]	['Agriculture', 'Labor', 'Missing', 'Professional']	['Crafts', 'Sales', 'Service']	1.073507656	4
[1.095559927326196, 1.0447060737947207]	['Agriculture', 'Missing', 'Professional', 'Service']	['Crafts', 'Labor', 'Sales']	1.072909621	4
[1.0927177278350615, 1.067836320199731]	['Labor', 'Missing', 'Professional', 'Sales']	['Agriculture', 'Crafts', 'Service']	1.074225866	4
[1.085804291396592, 1.0702661113667726]	['Missing', 'Professional', 'Sales', 'Service']	['Agriculture', 'Crafts', 'Labor']	1.07446142	4
[1.1204722327452645, 1.0308769235107282]	['Labor', 'Missing', 'Professional', 'Service']	['Agriculture', 'Crafts', 'Sales']	1.072072847	4
[1.01333671386026, 1.0953977462439153]	['Agriculture', 'Crafts', 'Professional', 'Sales']	['Labor', 'Missing', 'Service']	1.073717221	4
[1.0815273656592308, 1.0667719532421294]	['Agriculture', 'Crafts', 'Labor', 'Professional']	['Missing', 'Sales', 'Service']	1.07347091	4
[1.0764816952177274, 1.0691513112283386]	['Agriculture', 'Crafts', 'Professional', 'Service']	['Labor', 'Missing', 'Sales']	1.072576067	4
[1.0410218302426482, 1.0811365804806505]	['Crafts', 'Labor', 'Professional', 'Sales']	['Agriculture', 'Missing', 'Service']	1.074341142	4
[1.0325916220970428, 1.0833968299697654]	['Crafts', 'Professional', 'Sales', 'Service']	['Agriculture', 'Labor', 'Missing']	1.074119799	4
[1.1027261973336537, 1.0539185960148778]	['Crafts', 'Labor', 'Professional', 'Service']	['Agriculture', 'Missing', 'Sales']	1.072094547	4
[1.0790146722738436, 1.0727065649901748]	['Agriculture', 'Labor', 'Professional', 'Sales']	['Crafts', 'Missing', 'Service']	1.074077948	4
[1.072250544494806, 1.075124192156846]	['Agriculture', 'Professional', 'Sales', 'Service']	['Crafts', 'Labor', 'Missing']	1.074461529	4
[1.1168839156768722, 1.0406431419295887]	['Agriculture', 'Labor', 'Professional', 'Service']	['Crafts', 'Missing', 'Sales']	1.072694763	4
[1.141696498492564, 1.06336735070003]	['Labor', 'Professional', 'Sales', 'Service']	['Agriculture', 'Crafts', 'Missing']	1.074004449	4
[1.0335086018380089, 1.1171463666766375]	['Agriculture', 'Crafts', 'Missing', 'Sales']	['Labor', 'Professional', 'Service']	1.072216159	4
[1.0722954900062025, 1.0803059719032344]	['Agriculture', 'Crafts', 'Labor', 'Missing']	['Professional', 'Sales', 'Service']	1.074482352	4
[1.069827265560206, 1.0869271931118374]	['Agriculture', 'Crafts', 'Missing', 'Service']	['Labor', 'Professional', 'Sales']	1.074269827	4
[1.0479383445566497, 1.0928477755340176]	['Crafts', 'Labor', 'Missing', 'Sales']	['Agriculture', 'Professional', 'Service']	1.072979843	4
[1.0455566089719206, 1.097060279151978]	['Crafts', 'Missing', 'Sales', 'Service']	['Agriculture', 'Labor', 'Professional']	1.073595207	4
[1.0846330429881488, 1.0552276269317562]	['Crafts', 'Labor', 'Missing', 'Service']	['Agriculture', 'Professional', 'Sales']	1.074205882	4
[1.0637767567882996, 1.0801497530150346]	['Agriculture', 'Labor', 'Missing', 'Sales']	['Crafts', 'Professional', 'Service']	1.072120436	4
[1.061528458852803, 1.0848685051635965]	['Agriculture', 'Missing', 'Sales', 'Service']	['Crafts', 'Labor', 'Professional']	1.073114458	4
[1.093727830044641, 1.029287031473372]	['Agriculture', 'Labor', 'Missing', 'Service']	['Crafts', 'Professional', 'Sales']	1.073970281	4
[1.0828457263946665, 1.0664289628782015]	['Labor', 'Missing', 'Sales', 'Service']	['Agriculture', 'Crafts', 'Professional']	1.073140136	4
[1.0353672141402663, 1.0973163471734606]	['Agriculture', 'Crafts', 'Labor', 'Sales']	['Missing', 'Professional', 'Service']	1.072350847	4
[1.0334150018958645, 1.1013635685496768]	['Agriculture', 'Crafts', 'Sales', 'Service']	['Labor', 'Missing', 'Professional']	1.073083375	4
[1.0795077429354574, 1.0665485706655906]	['Agriculture', 'Crafts', 'Labor', 'Service']	['Missing', 'Professional', 'Sales']	1.074401829	4
[1.053175602362247, 1.0832250496205886]	['Crafts', 'Labor', 'Sales', 'Service']	['Agriculture', 'Missing', 'Professional']	1.073567157	4
[1.0732636289741269, 1.0724682552792126]	['Agriculture', 'Labor', 'Sales', 'Service']	['Crafts', 'Missing', 'Professional']	1.072762066	4
[1.0186772027345765, 1.0933546901350217]	['Agriculture', 'Crafts', 'Missing', 'Professional', 'Sales']	['Labor', 'Service']	1.07384893	5
[1.084832738893362, 1.0640981251431965]	['Agriculture', 'Crafts', 'Labor', 'Missing', 'Professional']	['Sales', 'Service']	1.073449436	5
[1.0796798485518313, 1.0664137382279368]	['Agriculture', 'Crafts', 'Missing', 'Professional', 'Service']	['Labor', 'Sales']	1.072571867	5
[1.0496159707696378, 1.079346323707148]	['Crafts', 'Labor', 'Missing', 'Professional', 'Sales']	['Agriculture', 'Service']	1.074399193	5
[1.0405268873227713, 1.0815760236756207]	['Crafts', 'Missing', 'Professional', 'Sales', 'Service']	['Agriculture', 'Labor']	1.074203599	5
[1.106818019118475, 1.0516215330913017]	['Crafts', 'Labor', 'Missing', 'Professional', 'Service']	['Agriculture', 'Sales']	1.072011115	5
[1.0858785470267547, 1.0708182452602073]	['Agriculture', 'Labor', 'Missing', 'Professional', 'Sales']	['Crafts', 'Service']	1.074047174	5
[1.0787178435864888, 1.0732011608648113]	['Agriculture', 'Missing', 'Professional', 'Sales', 'Service']	['Crafts', 'Labor']	1.074456758	5
[1.1205464091449882, 1.0381822330626997]	['Agriculture', 'Labor', 'Missing', 'Professional', 'Service']	['Crafts', 'Sales']	1.07256104	5
[1.1533047155511529, 1.0616717625302443]	['Labor', 'Missing', 'Professional', 'Sales', 'Service']	['Agriculture', 'Crafts']	1.073840619	5
[1.011181396991377, 1.0831687384502102]	['Agriculture', 'Crafts', 'Labor', 'Professional', 'Sales']	['Missing', 'Service']	1.074026346	5
[1.0041174226279344, 1.085382764827361]	['Agriculture', 'Crafts', 'Professional', 'Sales', 'Service']	['Labor', 'Missing']	1.073989364	5
[1.1005520518185585, 1.0587133975770677]	['Agriculture', 'Crafts', 'Labor', 'Professional', 'Service']	['Missing', 'Sales']	1.072520153	5
[1.0848976516491609, 1.0738491304787368]	['Crafts', 'Labor', 'Professional', 'Sales', 'Service']	['Agriculture', 'Missing']	1.074350733	5
[1.1487127055891695, 1.0662013859582715]	['Agriculture', 'Labor', 'Professional', 'Sales', 'Service']	['Crafts', 'Missing']	1.073907943	5
[1.0388985423848909, 1.0947928997116794]	['Agriculture', 'Crafts', 'Labor', 'Missing', 'Sales']	['Professional', 'Service']	1.072435157	5
[1.0368434189492128, 1.098775860703346]	['Agriculture', 'Crafts', 'Missing', 'Sales', 'Service']	['Labor', 'Professional']	1.073185376	5
[1.0819741455142986, 1.0628364325882558]	['Agriculture', 'Crafts', 'Labor', 'Missing', 'Service']	['Professional', 'Sales']	1.074376473	5
[1.057696034205267, 1.0810283932597915]	['Crafts', 'Labor', 'Missing', 'Sales', 'Service']	['Agriculture', 'Professional']	1.07359937	5
[1.0772500145098465, 1.0701243089471246]	['Agriculture', 'Labor', 'Missing', 'Sales', 'Service']	['Crafts', 'Professional']	1.072735167	5
[1.040929686432672, 1.0854689598987637]	['Agriculture', 'Crafts', 'Labor', 'Sales', 'Service']	['Missing', 'Professional']	1.073042503	5
[1.0223048390077045, 1.08146199500917]	['Agriculture', 'Crafts', 'Labor', 'Missing', 'Professional', 'Sales']	['Service']	1.074126508	6
[1.0141965601531442, 1.083648134315872]	['Agriculture', 'Crafts', 'Missing', 'Professional', 'Sales', 'Service']	['Labor']	1.074119378	6
[1.105165681538302, 1.0565487841441326]	['Agriculture', 'Crafts', 'Labor', 'Missing', 'Professional', 'Service']	['Sales']	1.07244651	6
[1.1194267388242052, 1.0723009830941819]	['Crafts', 'Labor', 'Missing', 'Professional', 'Sales', 'Service']	['Agriculture']	1.074299115	6
[1.1656875625039393, 1.064580524786143]	['Agriculture', 'Labor', 'Missing', 'Professional', 'Sales', 'Service']	['Crafts']	1.073720601	6
[0.35335933502142136, 1.0759755834161266]	['Agriculture', 'Crafts', 'Labor', 'Professional', 'Sales', 'Service']	['Missing']	1.073807735	6
[1.0460584911356179, 1.0833972877916547]	['Agriculture', 'Crafts', 'Labor', 'Missing', 'Sales', 'Service']	['Professional']	1.07309178	6

g) (5 points). What is the optimal split for the JobCategory predictor?

Optimal Split entropy for Job: 1.0720111150297396 since it has the lowest entropy.

['Agriculture', 'Sales'] and ['Crafts', 'Labor', 'Missing', 'Professional', 'Service']

h) (5 points). Between the CreditCard and the JobCategory predictors, which predictor will you choose for producing the second layer (i.e., depth 1) of your decision tree?

Split Entropy=1.0708382285522746

We will choose the predictor Creditcard or producing the second layer of the tree ['Others', 'Visa'] and ['American Express', 'Discover', 'MasterCard'] as the entropy is lowest from both the predictors.

Question 2 (50 points)

In 2014, Allstate provided the data on Kaggle.com for the Allstate Purchase Prediction Challenge which is open. The data contain transaction history for customers that ended up purchasing a policy. For each Customer ID, you are given their quote history and the coverage options they purchased.

The data is available on the Blackboard as Purchase_Likelihood.csv. It contains 665,249 observations on 97,009 unique Customer ID. We are going to use the MNLogit function to build a multinomial logistic model to predict purchase likelihood of coverage A using three predictors. The target variable is **A** which have these categories 0, 1, and 2. The nominal predictors are (categories are inside the parentheses):

1. **group_size**. How many people will be covered under the policy (1, 2, 3 or 4)?
2. **homeowner**. Whether the customer owns a home or not (0=no, 1=yes)
3. **married_couple**. Does the customer group contain a married couple (0=no, 1=yes)

Please build a multinomial logistic model using and answer the following questions.

a) (2 points) Suppose you start with a model with only the Intercept term (i.e., without any predictors). How many parameters are in this model?
For intercept only model we only have one parameter is involved

b) (3 points) What are the marginal counts of the categories of the target variable A?

ToatlCount_0	ToatlCount_1	ToatlCount_2	ToatlCount
143691	426067	95491	665249

c) (5 points) Without calling the MNLogit function, what are the maximum likelihood estimates of the predicted probabilities $\pi_{ij}, j = 1,2,3$ of this Intercept-only model? Show all the necessary steps and the estimates for the $\pi_{ij}, j = 1,2,3$. (Hint: equate the first derivatives of the log-likelihood function to zeros for this Intercept-only model)

Total Observation	MLE	Odds	MLE Intercept
143691.0	0.215996	1.000000	0.000000
426067.0	0.640462	2.965161	1.086931
95491.0	0.143542	0.664558	-0.408633

d) (3 points) What is the log-likelihood value of this Intercept-only model? (Hint: the log-likelihood function is $l = \sum_{i=1}^{m} \sum_{j=1}^{3} n_{ij} \log_e(\pi_{ij})$)

(tototalCount_0)log$_e$(prob_0) + (tototalCount_1)log$_e$(prob_1) + (tototalCount_2)log$_e$(prob_2)
143691 log$_e$(0.215996) + 426067 log$_e$(0.640462) + 95491 log$_e$(0.143542)
Log Likelihood: -595406.7618844224

e) (5 points) Next, you are asked to mathematically calculate the maximum likelihood estimates of the Intercept terms $\beta_{j0}, j = 1,2,3$. The convention is to set the Intercept term to zero for the target category A = 0, i.e., $\beta_{10} = 0$. (Hint: use the mathematical formula of the logit of π_{ij} (i.e., $\log_e(\pi_{ij}/\pi_{i1})$ for this Intercept only model, then solve for the betas)?

J = 1, 2, 3
β = [β10, β20, β30]

$\pi i1$ = Probability of 0= 0.21599581510081187
$\pi i2$ = Probability of 1= 0.64046244338586
$\pi i3$ = Probability of 2= 0.1435417415133281

βj =loge ($\pi ij/\pi iJ$)
β 1 = loge ($\pi i0/\pi i0$) = loge(1) = 0
β 2 = loge ($\pi i1/\pi i0$) = loge(0.6404624433858/0.2159958151008) = 1.08693
β 3 = loge ($\pi i3/\pi i1$) = loge(0. 143541741513/0. 2159958151008) = -0.40863

f) (5 points) Create and display a contingency table where group_size, homeowner, and married_couple are on the row dimension, and A is on the column dimension. The cell contents are the row percentages of the categories of A per each level combination of group_size, homeowner, and married_couple.

group_size	homeowner	married_couple	0	1	2	All
1	0	0	25.77842	59.14609	15.07549	100
1	0	1	32.14196	51.69897	16.15907	100
1	1	0	18.02423	68.62947	13.34631	100
1	1	1	22.17153	62.39966	15.42881	100
2	0	0	27.27079	55.59455	17.13465	100
2	0	1	20.55084	64.5793	14.86986	100
2	1	0	25.33176	59.48539	15.18285	100
2	1	1	16.06721	70.20728	13.72551	100
3	0	0	27.55906	66.92913	5.511811	100
3	0	1	23.30023	59.51108	17.18869	100
3	1	0	25.99532	59.48478	14.51991	100
3	1	1	26.15788	56.30644	17.53568	100
4	0	0	100	0	0	100
4	0	1	18.75	67.85714	13.39286	100
4	1	0	48.48485	42.42424	9.090909	100
4	1	1	33.33333	53.00546	13.6612	100
All			21.59958	64.04624	14.35417	100

g) (2 points) Based on the contingency table in e), do you expect the separation or the quasi-separation phenomenon to occur when we build the multinomial logistic model which has group_size, homeowner, and married_couple as the predictors.
Based on the contingency table in (f), we expect the Quasi-complete separation because none of the predictor completely define particular category of target variable.

h) (5 points) Now, you will use the MNLogit function to build the multinomial logistic model which has group_size, homeowner, and married_couple as the predictors. What value of the target variable A is used by the MNLogit function as the reference category? Next, what is the log-likelihood value of this model? Finally, how many parameters (including the redundant ones) are in the model?
Reference Category Value of A=0
Log-Likelihood=-591936.7938327907
Since there are 9 parameters and each parameter has its occurrence and nonoccurrence i.e. So,Parameters Involved=9*2=18

i) (10 points) What are the values of group_size, homeowner, and married_couple such that the odd Prob(A=1)/Prob(A = 0) will attain its maximum? What is the maximum odd Prob(A = 1)/Prob(A = 0) value?
Predicted probability of 0: 0.1606721019540383
Predicted probability of 1: 0.7020727585496748
Maximum odd Probability: 4.36959963809093

j) (5 points) According to the multinomial logistic model, what is the odds ratio for group_size = 3 versus group_size = 1, and A = 2 versus A = 0? Mathematically, the odds ratio is (Prob(A=2)/Prob(A=0) | group_size = 3) / ((Prob(A=2)/Prob(A=0) | group_size = 1).
group_size_3A2=-0.107280
group_size_1A0=0.411100
group_size_3A2-group_size_1A0
Required Odds Ratio: 0.5954844518092098

k) (5 points) According to the multinomial logistic model, what is the odds ratio for group_size = 1 versus group_size = 3, and A = 2 versus A = 1? Mathematically, the odds ratio is (Prob(A=2)/Prob(A=1) | group_size = 1) / ((Prob(A=2)/Prob(A=1) | group_size = 3).

group_size_3A2=-0.026011
group_size_1A1=0.108286
group_size_1A1-group_size_3A2
Required Odds Ratio: 1.1437324577458428