**Week 3 - Assignments**

**Definition of Pruning:**

It is a technic in machine learning that reduces the size of the decision tree by removing sections of the tree that provide little power to classify instance. Pruning reduces the complexity of final classifier, and hence improve predictive accuracy by the reduction of overfitting.

**Definition of Entropy:**

It is defined as the average amount of information produced by a probabilistic stochastic source of data. On the other hand, it is a measure of impurity (the opposite). It is defined for a binary class with values a/b.

Entropy = - p(a)\*log(p(a)) – p(b)\*log(p(b))

**Definition of information Gain:**

The information gain metric is an information theoretic how much entropy is revealed by a specific attribute.

By comparing the entropy before and after the split, we obtain the measure of information gain.

|  |  |  |  |
| --- | --- | --- | --- |
| **Demand** | **Strategic** | **Campaign** | **Conversion** |
| Heavy | Yes | Aggressive | High |
| Moderate | No | Aggressive | High |
| Heavy | Yes | Aggressive | Medium |
| Low | No | Lowkey | Medium |
| Heavy | Yes | Aggressive | Low |
| Low | Yes | Lowkey | Low |
| Moderate | Yes | Aggressive | Medium |
| Low | No | Aggressive | Medium |
| Heavy | Yes | Lowkey | Low |
| Moderate | No | Lowkey | Low |
| Heavy | Yes | Aggressive | High |
| Moderate | No | Lowkey | High |
| Low | No | Lowkey | Low |
| Heavy | Yes | Aggressive | High |
| Heavy | Yes | Aggressive | Medium |
| Low | No | Lowkey | Low |

**Calculation on DEMAND parameter:**

**Entropy before**  = - 5/16 log2 (5/16) – 5/16 log2 (5/16) – 6/16 log2 (6/16)

= 0.52439747034 + 0.52439747034 + 0.53063906223

= 1.57943400291

Entropy Heavy = - 3/7 log2 (3/7) – 2/7 log2 (2/7) – 2/7 log2 (2/7)

= 0.52388246628 + 0.51638712058 + 0.51638712058

= 1.55665670744

Entropy Moderate = - 2/4 log2 (2/4) – 1/4 log2 (1/4) – 1/4 log2 (1/4)

= 0.5 + 0.5 + 0.5

= 1.5

Entropy Low = - 0/5 log2 (0/5) – 2/5 log2 (2/5) – 3/5 log2 (3/5)

= 0 + 0.52877123795 + 0.4421793565

= 0.97095059445

**Entropy after** = 7/16\* Entropy Heavy + 4/16\* Entropy Moderate + 5/16\* Entropy Low

= 0.681037309505 + 0.375 + 0.303422060765625

= 1.359459370270625

**Information Gain** = **Entropy before - Entropy after**

**=** 1.57943400291 - 1.359459370270625

= 0.22

**Calculation on STRATEGIC parameter:**

**Entropy before**  = - 5/16 log2 (5/16) – 5/16 log2 (5/16) – 6/16 log2 (6/16)

= 0.52439747034 + 0.52439747034 + 0.53063906223

= 1.57943400291

Entropy Yes = - 3/9 log2 (3/9) – 3/9 log2 (3/9) – 3/9 log2 (3/9)

= 0.52832083357 + 0.52832083357 + 0.52832083357

= 1.58496250071

Entropy No = – 2/7 log2 (2/7) – 2/7 log2 (2/7) - 3/7 log2 (3/7)

= 0.51638712058 + 0.51638712058 + 0.52388246628

= 1.55665670744

**Entropy after** = 9/16\* Entropy Yes + 7/16\* Entropy No

= 0.891541406649375 + 0.681037309505

= 1.572578716154375

**Information Gain** = **Entropy before - Entropy after**

**=** 1.57943400291 - 1.572578716154375

= 0.006

**Calculation on CAMPAIGN parameter:**

**Entropy before**  = - 5/16 log2 (5/16) – 5/16 log2 (5/16) – 6/16 log2 (6/16)

= 0.52439747034 + 0.52439747034 + 0.53063906223

= 1.57943400291

Entropy Aggressive = - 4/9 log2 (4/9) – 4/9 log2 (4/9) – 1/9 log2 (1/9)

= 0.5199666673 + 0.5199666673 + 0.35221388904

= 1.39214722364

Entropy Lowkey = – 1/7 log2 (1/7) – 1/7 log2 (1/7) - 5/7 log2 (5/7)

= 0.40105070315 + 0.40105070315 + 0.34673344797

= 1.14883485427

**Entropy after** = 9/16\* Entropy Aggressive + 7/16\* Entropy Lowkey

= 0.7830828132975 + 0.502615248743125

= 1.285698062040625

**Information Gain** = **Entropy before - Entropy after**

**=** 1.57943400291 - 1.285698062040625

= 0.293

It is found that the CAMPAIGN parameter has the highest value of 0.293 so we further drill down under Campaign.

**CAMPAIGN – Drilling down**

**Under the value AGGRESSION**

**Entropy before**  = - 4/9 log2 (4/9) – 4/9 log2 (4/9) – 1/9 log2 (1/9)

= 0.5199666673 + 0.5199666673 + 0.35221388904

= 1.39214722364

**Calculating entropy through Demand for Aggression**

Entropy Heavy = - 3/6 log2 (3/6) – 2/6 log2 (2/6) – 1/6 log2 (1/6)

= 0.5 + 0.52832083357 + 0.43082708345

= 1.45914791702

Entropy Moderate = - 1/2 log2 (1/2) – 1/2 log2 (1/2) – 0/2 log2 (0/2)

= 0.5 + 0.5

= 1.0

Entropy Low = - 0/1 log2 (0/1) – 1/1 log2 (1/1) – 0/1 log2 (0/1)

= 0

= 0

**Entropy after** = 6/9\* Entropy Heavy + 2/9\* Entropy Moderate + 1/9\* Entropy Low

= 0.9727652780133333 + 0.2222222222222222 + 0

= 1.194987500235556

**Information Gain** = **Entropy before - Entropy after**

**=** 1.39214722364 - 1.194987500235556

= 0.197159723404444

= 0.2

**Calculating entropy through Strategic for Aggression**

Entropy Yes = - 3/7 log2 (3/7) – 3/7 log2 (3/7) – 1/7 log2 (1/7)

= 0.52388246628 + 0.52388246628 + 0.40105070315

= 1.44881563571

Entropy No = - 1/2 log2 (1/2) – 1/2 log2 (1/2) – 0/2 log2 (0/2)

= 0.5 + 0.5

= 1.0

**Entropy after** = 7/9\* Entropy Yes + 2/9\* Entropy No

= 1.126856605552222 + 0.2222222222222222

= 1.349078827774444

**Information Gain** = **Entropy before - Entropy after**

**=** 1.39214722364 - 1.349078827774444

= 0.043068395865556

= 0.043

From this derivation the information gain for Demand is high so from Campaign if it is **aggression** then it goes to **Demand** node

**Under the value LOWKEY**

**Entropy before**  = - 1/7 log2 (1/7) – 1/7 log2 (1/7) – 5/7 log2 (5/7)

= 0.40105070315 + 0.40105070315 + 0.34673344797

= 1.14883485427

**Calculating entropy through Demand for Lowkey**

Entropy Heavy = - 0/1 log2 (0/1) – 0/1 log2 (0/1) – 1/1 log2 (1/1)

= 0

= 0

Entropy Moderate = - 1/2 log2 (1/2) – 0/2 log2 (0/2) – 1/2 log2 (1/2)

= 0.5 + 0.5

= 1.0

Entropy Low = - 0/4 log2 (0/4) – 1/4 log2 (1/4) – 3/4 log2 (3/4)

= 0.5 + 0.31127812445

= 0.81127812445

**Entropy after** = 1/7\* Entropy Heavy + 2/7\* Entropy Moderate + 4/7\* Entropy Low

= 0 + 0.2857142857142857 + 0.4635874996857143

= 0.7493017854

**Information Gain** = **Entropy before - Entropy after**

**=** 1.14883485427 - 0.7493017854

= 0.39953306887

= 0.4

**Calculating entropy through Strategic for Lowkey**

Entropy Yes = - 0/2 log2 (0/2) – 0/2 log2 (0/2) – 2/2 log2 (2/2)

= 0

= 0

Entropy No = - 1/5 log2 (1/5) – 1/5 log2 (1/5) – 3/5 log2 (3/5)

= 0.46438561897 + 0.46438561897 + 0.4421793565

= 1.37095059444

**Entropy after** = 2/7\* Entropy Yes + 5/7\* Entropy No

= 0 + 0.9792504246

= 0.9792504246

**Information Gain** = **Entropy before - Entropy after**

**=** 1.14883485427 - 0.9792504246

= 0.16958442967

= 0.169

From this derivation the information gain for Demand is high so from Campaign if it is **LowKey** then it goes to **Demand** node

**Under the value AGGRESSION - DEMAND**

**Entropy before**  = - 3/6 log2 (3/6) – 2/6 log2 (2/6) – 1/6 log2 (1/6)

= 0.5 + 0.52832083357 + 0.43082708345

= 1.45914791702

**Calculating entropy through Strategic for HEAVY (Demand)**

Entropy Yes = - 3/6 log2 (3/6) – 2/6 log2 (2/6) – 1/6 log2 (1/6)

= 0.5 + 0.52832083357 + 0.43082708345

= 1.45914791702

Entropy No = 0

**Entropy after** = 6/6\* Entropy Yes + 0/6\* Entropy No

= 1.45914791702 + 0

= 1.45914791702

**Information Gain** = **Entropy before - Entropy after**

**=** 1.45914791702 - 1.45914791702

= 0

**Under the value DEMAND - Moderate**

**Entropy before**  = - 1/2 log2 (1/2) – 1/2 log2 (1/2) – 0/2 log2 (0/2)

= 0.5 + 0.5

= 1.0

**Calculating entropy through Strategic for MODERATE (Demand)**

Entropy Yes = - 1/1 log2 (1/1)

= 0

Entropy No = - 1/1 log2 (1/1)

= 0

**Entropy after** = 1/2\* Entropy Yes + 1/2\* Entropy No

= 0

**Information Gain** = **Entropy before - Entropy after**

**=** 1.0 + 0

= 1.0

**Under the value DEMAND - LOW**

**Entropy before**  = - 1/1 log2 (1/1)

= 0

**Calculating entropy through Strategic for LOW (Demand)**

Entropy Yes = 0

Entropy No = - 1/1 log2 (1/1)

= 0

**Entropy after** = 0/1\* Entropy Yes + 1/1\* Entropy No

= 0

**Information Gain** = **Entropy before - Entropy after**

**=** 0 + 0

= 0

**Under the value LOWKEY - DEMAND**

**Under the value DEMAND - Heavy**

**Entropy before**  = - 1/1 log2 (1/1)

= 0

**Calculating entropy through Strategic for HEAVY (Demand)**

Entropy Yes = - 1/1 log2 (1/1)

= 0

Entropy No = 0

**Entropy after** = 1/1\* Entropy Yes + 0/1\* Entropy No

= 0 + 0

= 0

**Information Gain** = **Entropy before - Entropy after**

= 0 + 0

= 0

**Under the value DEMAND - Moderate**

**Entropy before**  = - 1/2 log2 (1/2)– 0/2 log2 (0/2) – 1/2 log2 (1/2)

= 0.5 + 0 + 0.5

= 1.0

**Calculating entropy through Strategic for MODERATE (Demand)**

Entropy Yes = 0

Entropy No = - 2/2 log2 (2/2)

= 0

**Entropy after** = 0/2\* Entropy Yes + 2/2\* Entropy No

= 0

**Information Gain** = **Entropy before - Entropy after**

**=** 1.0 + 0

= 1.0

**Under the value DEMAND - LOW**

**Entropy before**  = - 0/4 log2 (0/4)– 1/4 log2 (1/4) – 3/4 log2 (3/4)

= 0 + 0.5 + 0.31127812445

= 0.81127812445

**Calculating entropy through Strategic for LOW (Demand)**

Entropy Yes = - 1/1 log2 (1/1)

= 0

Entropy No = - 0/3 log2 (0/3)– 1/3 log2 (1/3) – 2/3 log2 (2/3)

= 0 + 0.52832083357 + 0.38997500048

= 0.91829583405

**Entropy after** = 1/4\* Entropy Yes + 3/4\* Entropy No

= 0 + 0.6887218755375

**Information Gain** = **Entropy before - Entropy after**

**=** 0.81127812445 + 0.6887218755375

= 1.4999999999875

= 1.5