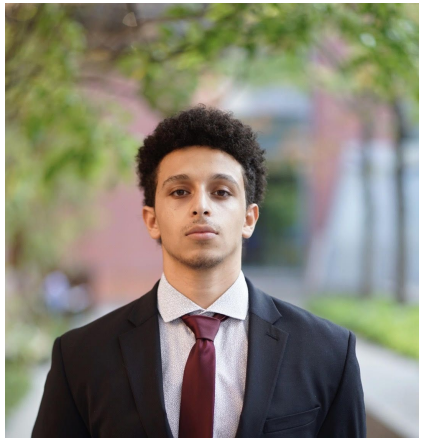
Comcast Customer Analysis

February 2021

Wharton Customer Analytics Accelerator

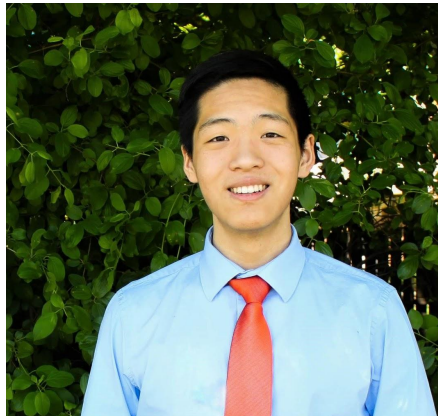
Wharton Customer Analytics Team

Cyrus Shanehsaz
Engagement Lead

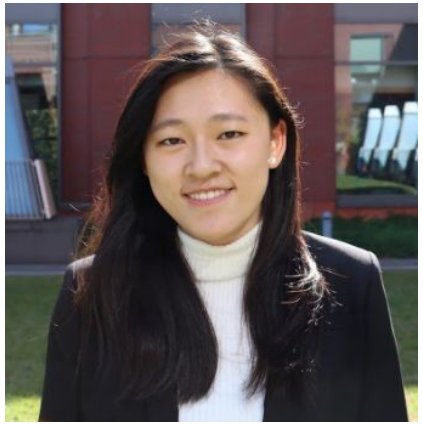


Ahmed Ahmed
Senior Analyst

Ashley Clarke
Senior Analyst



Matthew Dong
Junior Analyst



Emily Guo
Junior Analyst

Key Questions:

1. How do customers value individual products and bundles?
2. What regional and product segments of customers exist?
3. How do segments of customers value products differently?

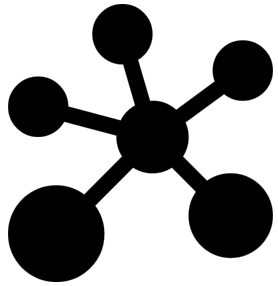
Agenda

1. Methodology
2. K-means Clustering & Random Forest
3. Pricing Model
4. Findings and Takeaways

Methodology

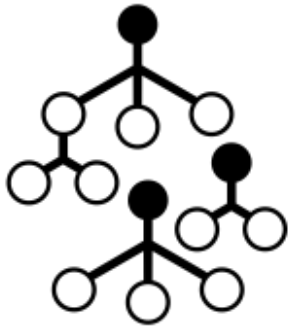
Methodology

Clustering



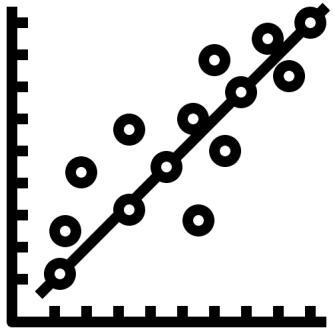
- Mathematical way to group customers
- Reveals pods of customers that emerge
- Compare with current customer segments to test validity

Random Forest



- More difficult to interpret
  - Challenges emerge when trying to make sense of the collection of trees
  - Less directly measurable coefficients
  - Shows range of coefficients instead
- Efficient
  - High prediction accuracy
  - Relatively fast training and prediction times
  - Scalable with the inclusion of additional data

Linear Modeling



- Highly interpretable
  - Shows direct values of each product
  - Easy to see effect of changing individual products in bundles
- Susceptible to overfitting
  - potential to overfit with too many collinear variables

Customer Segmentation with K-means Clustering

Methodology | Clustering and Random Forest | Pricing Model | Findings and Takeaways

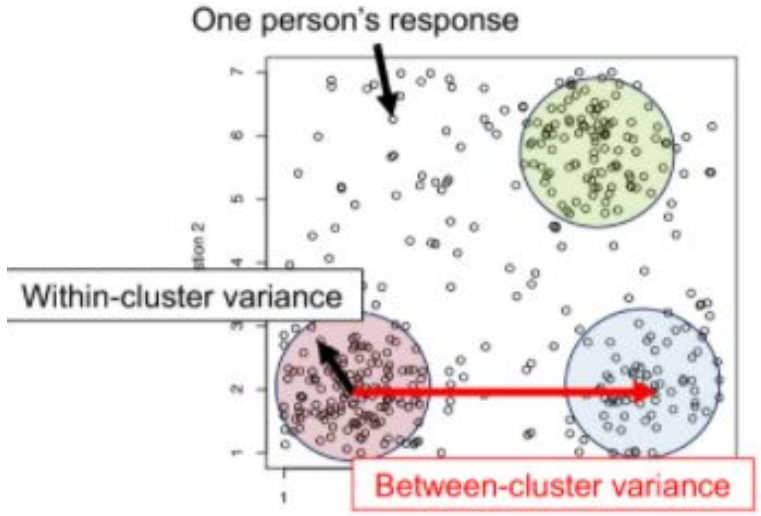
K-means Clustering

Why?

- Unsupervised machine learning technique: detects patterns in raw data without being fed labels
- Finds "natural" groupings between observations

What is it?

- An iterative technique which assigns each observation to the cluster closest to it
- Minimize within-cluster variance, maximize between-cluster variance



Clusters Overview

K-MEANS CLUSTERS FROM SAMPLE OF 2,000,000 CUSTOMERS FOR EFFICIENCY

(1) Video & Internet

- xx% of sample
- 93.3% Video/Internet customers
- Average customer revenue: $xxx

(2) Technologists

- xx% of sample
- 96.3% Internet Only customers
- Average customer revenue: $xxx

(3) Traditionalists

- xx% of sample
- 96% Video/Internet/Voice customers
- Average customer revenue: $xxx

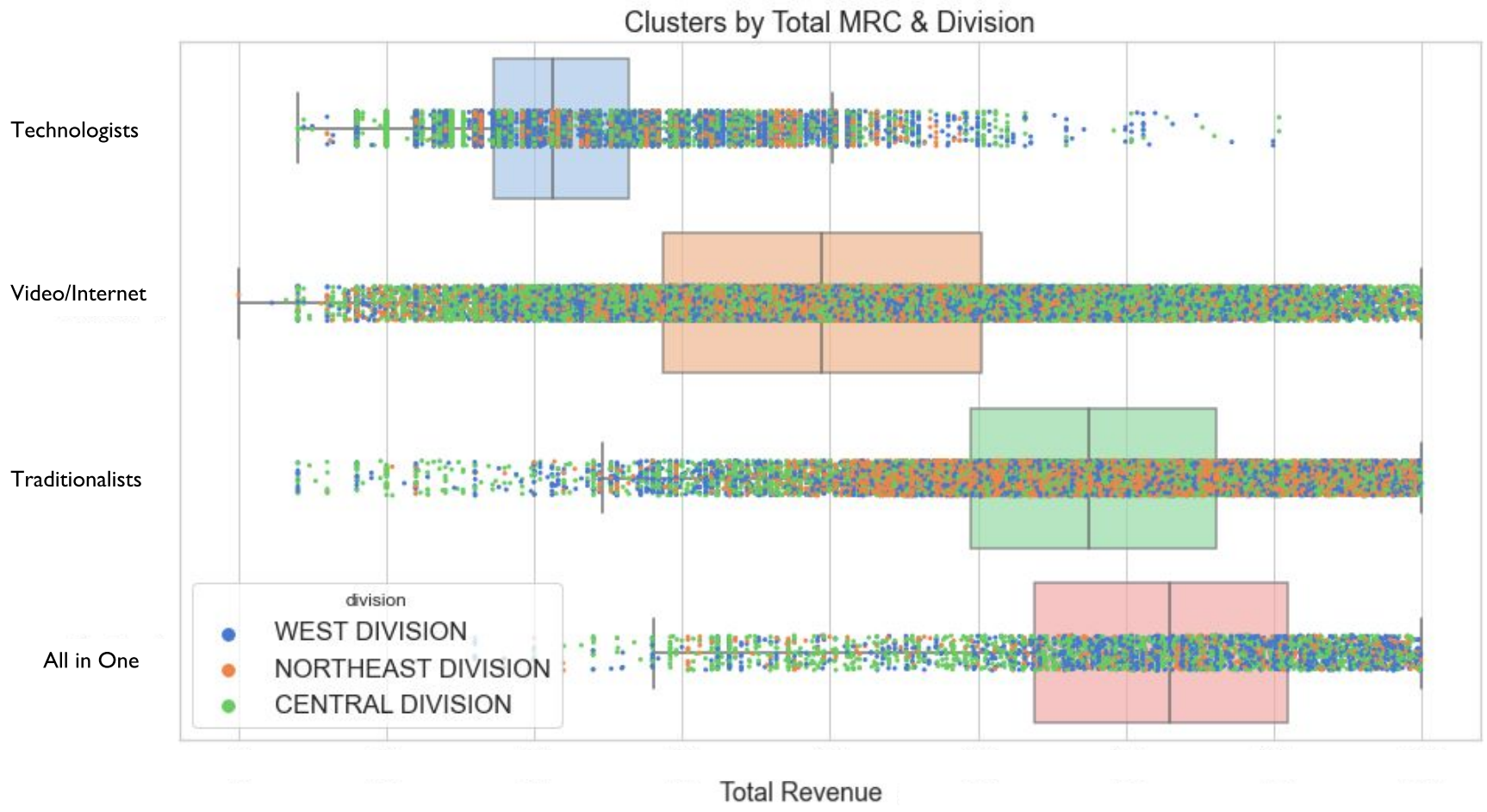
(4) All in One

- xx% of sample
- 90% Video/ Internet/ Voice/ Xfinity Home or Video/ Internet/ Xfinity Home customers
- Average customer revenue: $xxx

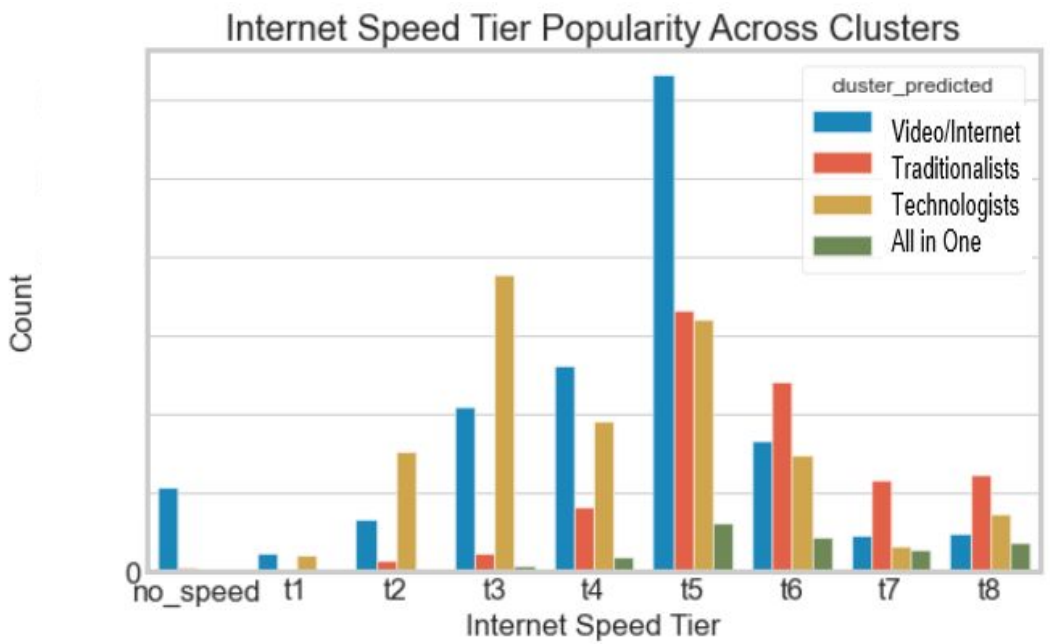
Clusters Overview – Revenue & Geography

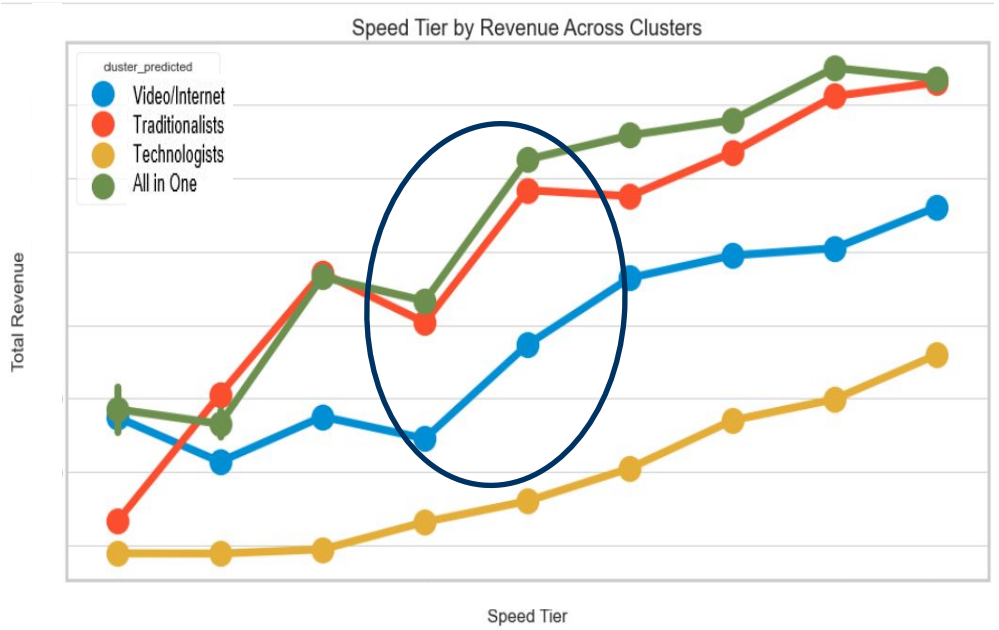
REVENUE AND DIVISIONAL DIFFERENCES

- Clusters show significant differences in revenue
- Regional affiliations exist in some clusters
  - xx% of Video/Internet Cluster customers are from the Central Division
  - xx% of Traditionalists are from the Northeast Division.
  - xx% of Technologists are from the West Division



Clusters Overview – Internet Speed Tiers





Internet Tier Popularity

- Video/Internet customers prefer t5 or below
- Technologists prefer moderate speeds (t3-t5)
- Traditionalists and All in One customers value high speeds (> t5)

Internet Tier by Total Revenue

- ¾ clusters willing to spend $xx-$xx more when jump from t3 to t4
- Technologists overall WTP is linearly correlated to speed tier

Exploring Customers' Values with Random Forests

Methodology | Clustering and Random Forest | Pricing Model | Findings and Takeaways

Random Forest

Advantages

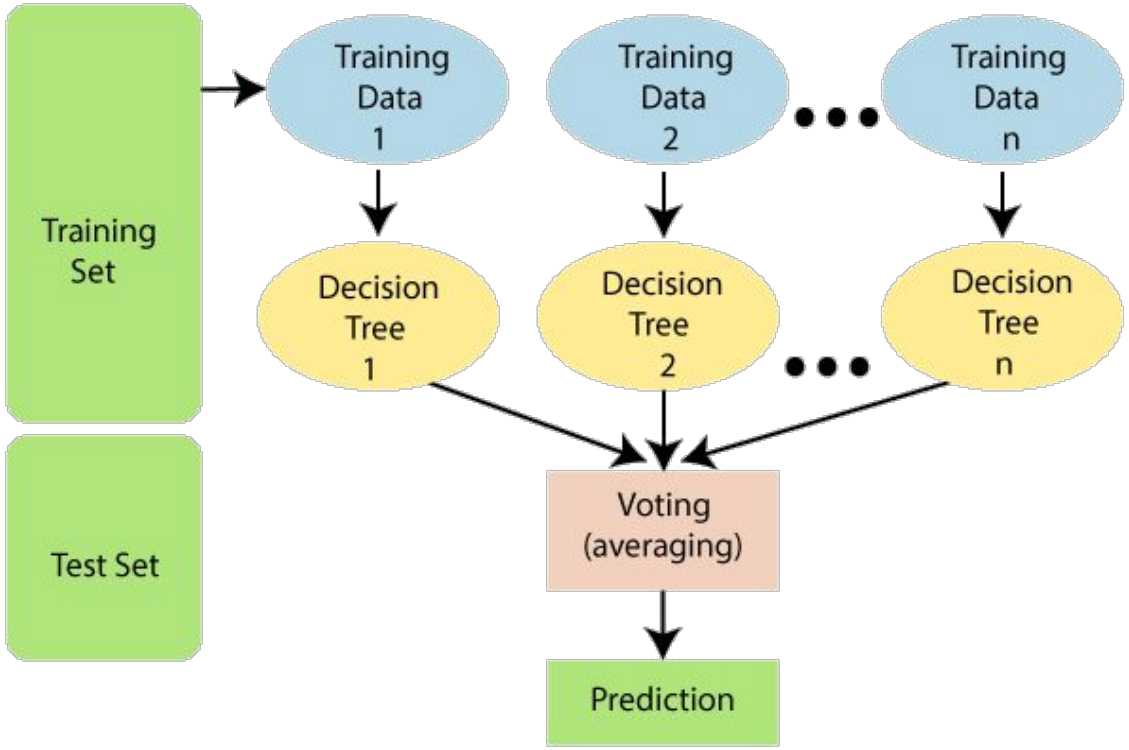
- Aggregated results of multiple individual decision trees
- Reduces variance
- Suitable for larger datasets

Disadvantages

- Higher difficulty in interpretation of results
- Longer training time

Method

- Data pre-processed to exclude business customers, then split by region
- Specific variables are binned, then encoded for improved interpretability
- 80% of data was used for training, 20% used for validation
- Number and depth of trees determined empirically



Random Forest Results (by Region)

The r^2 value represents the accuracy of the random forest model, and the coefficients below represent the importance of each feature in the regression prediction

Overall

r^2: 0.900717880474454

video_tier_name_new	0.68
promo_tier	0.13
speed	0.10
product_mix	0.05
competitor	0.02
hsd_tier_name_new	0.01
new_product	0.00
xh_tier_name_new	0.00
activity	0.00
cdv_tier_name_new	0.00

West

r^2: 0.9047519523347576

video_tier_name_new	0.67
speed	0.15
promo_tier	0.10
product_mix	0.05
competitor	0.01
hsd_tier_name_new	0.00
xh_tier_name_new	0.00
cdv_tier_name_new	0.00
new_product	0.00
activity	0.00

Central

r^2: 0.8730190024468925

promo_tier	0.65
video_tier_name_new	0.17
speed	0.08
competitor	0.04
product_mix	0.03
hsd_tier_name_new	0.01
new_product	0.00
xh_tier_name_new	0.00
activity	0.00
cdv_tier_name_new	0.00

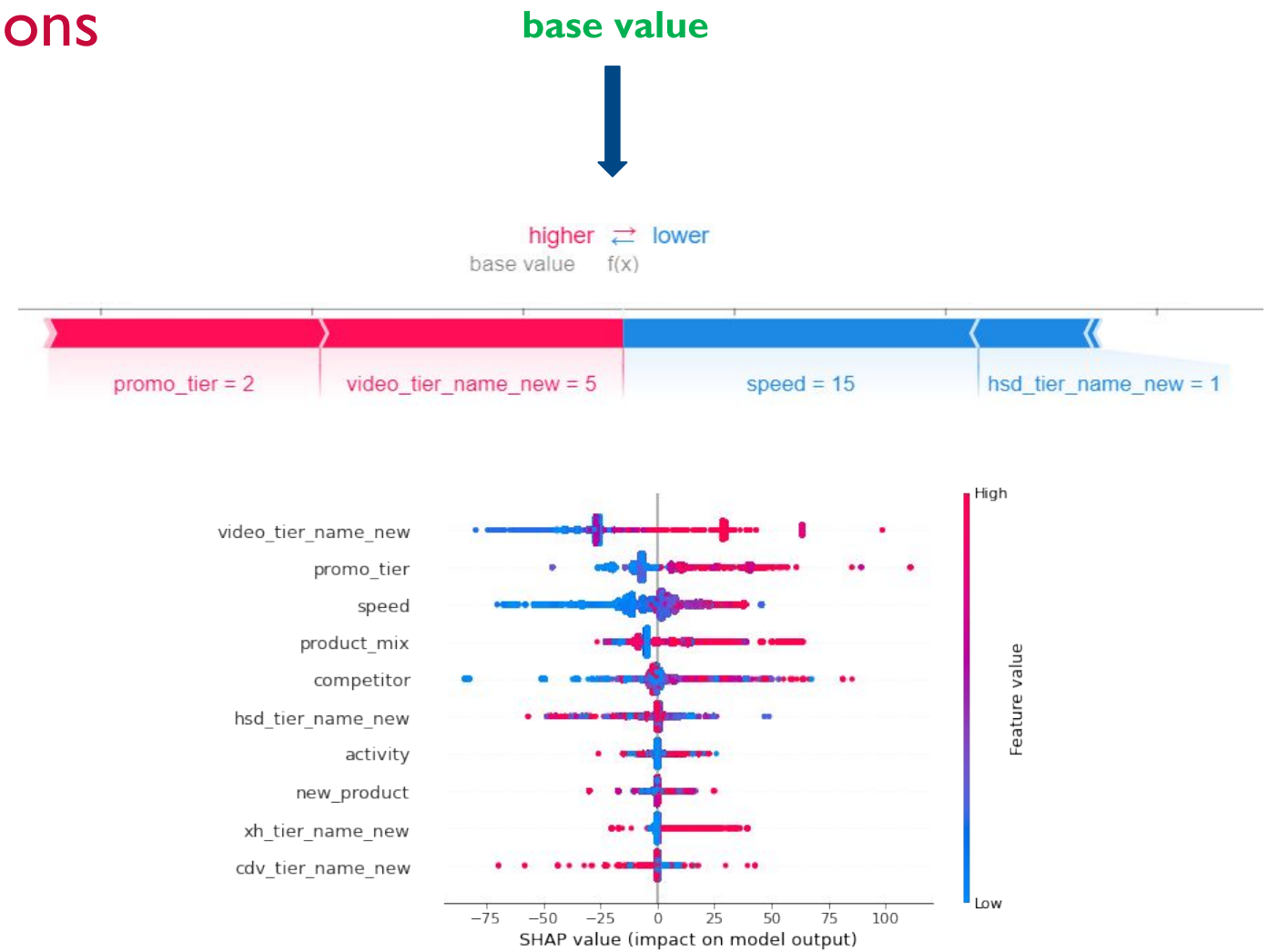
Northeast

r^2: 0.8964587937162071

video_tier_name_new	0.80
speed	0.07
promo_tier	0.05
product_mix	0.05
new_product	0.02
hsd_tier_name_new	0.01
xh_tier_name_new	0.00
activity	0.00
competitor	0.00
cdv_tier_name_new	0.00

SHAP: Shapely Additive Explanations

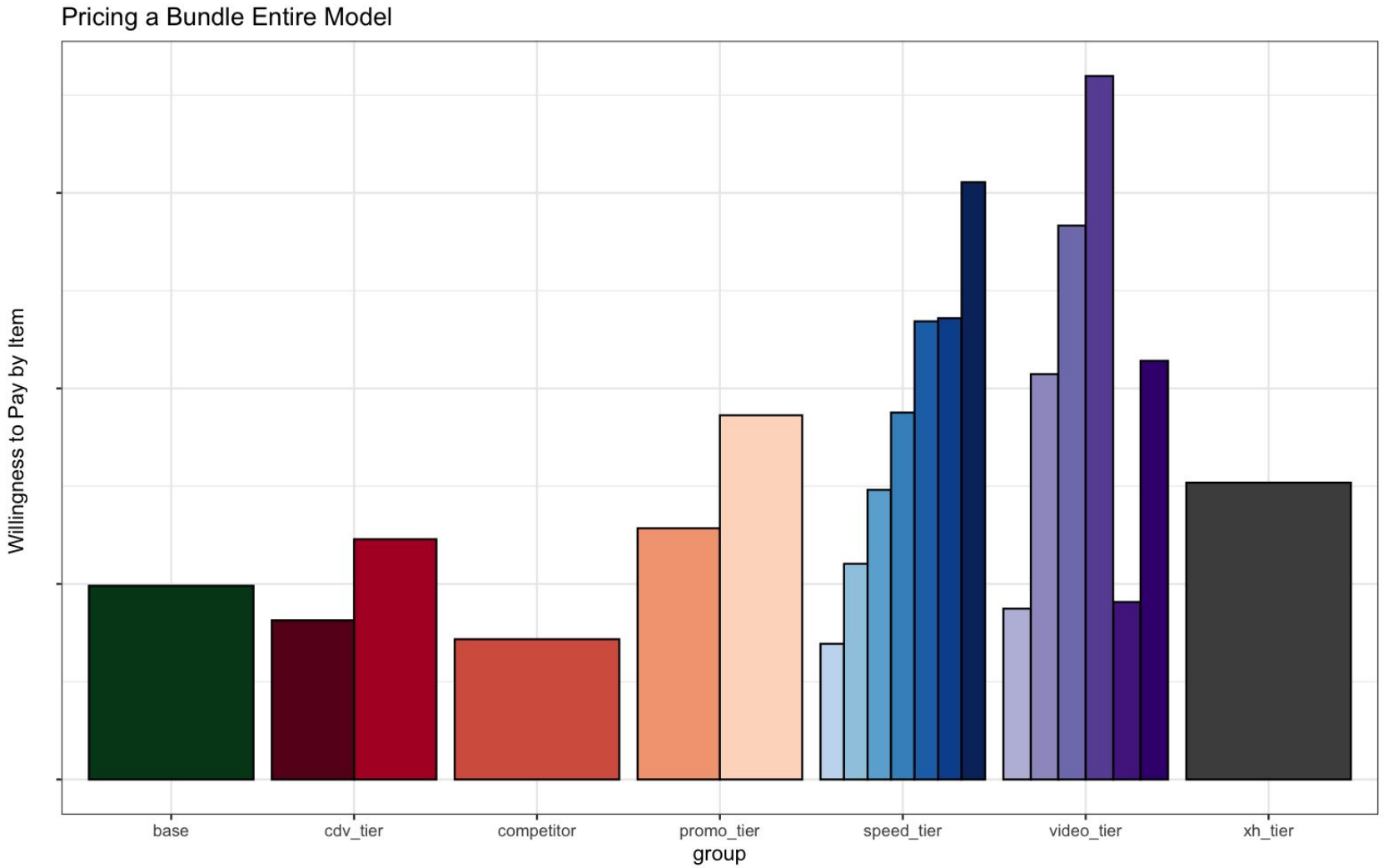
- The **force plot,** a visualization of SHAP
  - The **base value** is the average prediction of all data entries
  - Different values either increase or decrease the predicted value
  - Force plot visualizes how the value of a certain feature "pushes" the **predicted value f(x)** away from the base value by a certain amount: the SHAP value



Estimating Customers' Values With Linear Models

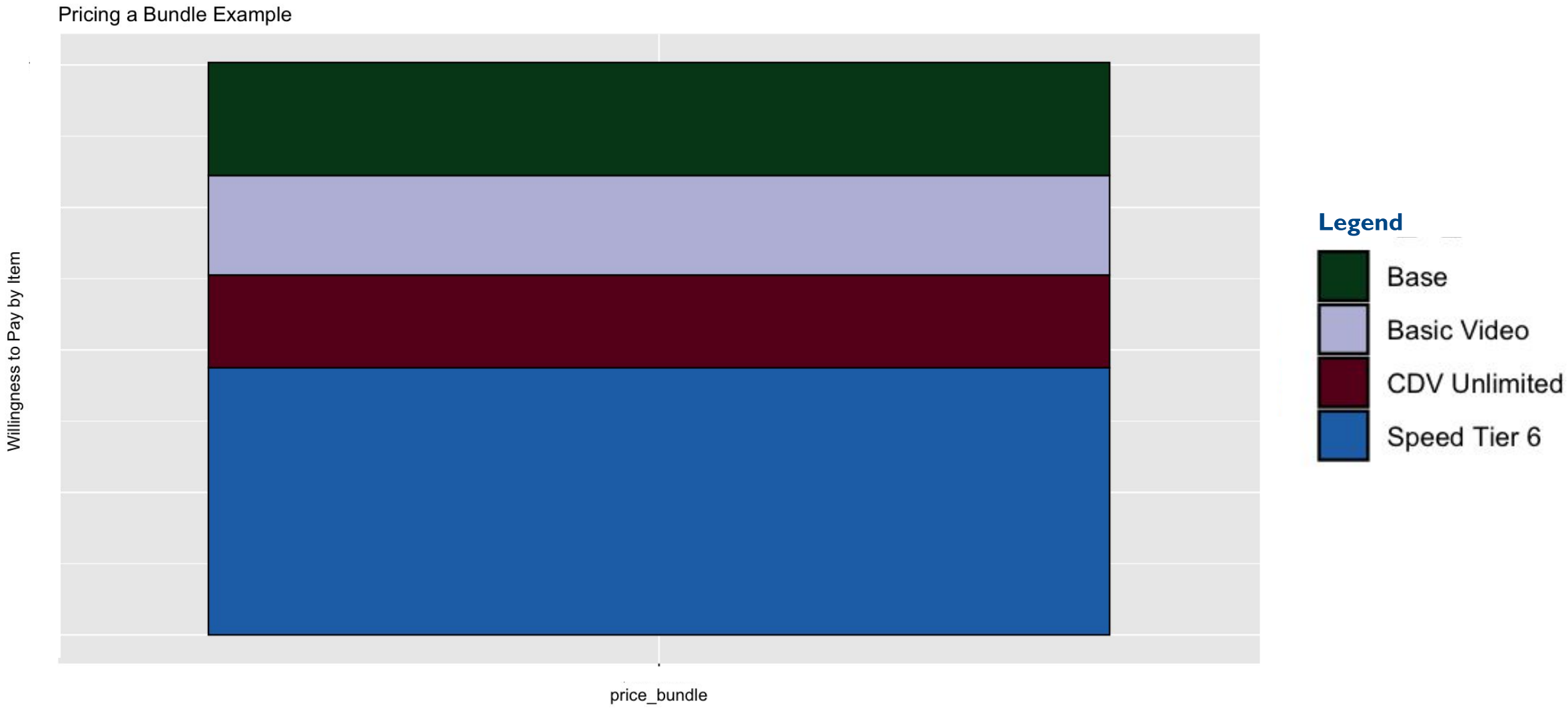
Methodology | Clustering and Random Forest | Pricing Model | Findings and Takeaways

Pricing Model All Regions (Values Greater than $10)



Legend

Example of How to Price Using All Regions Model

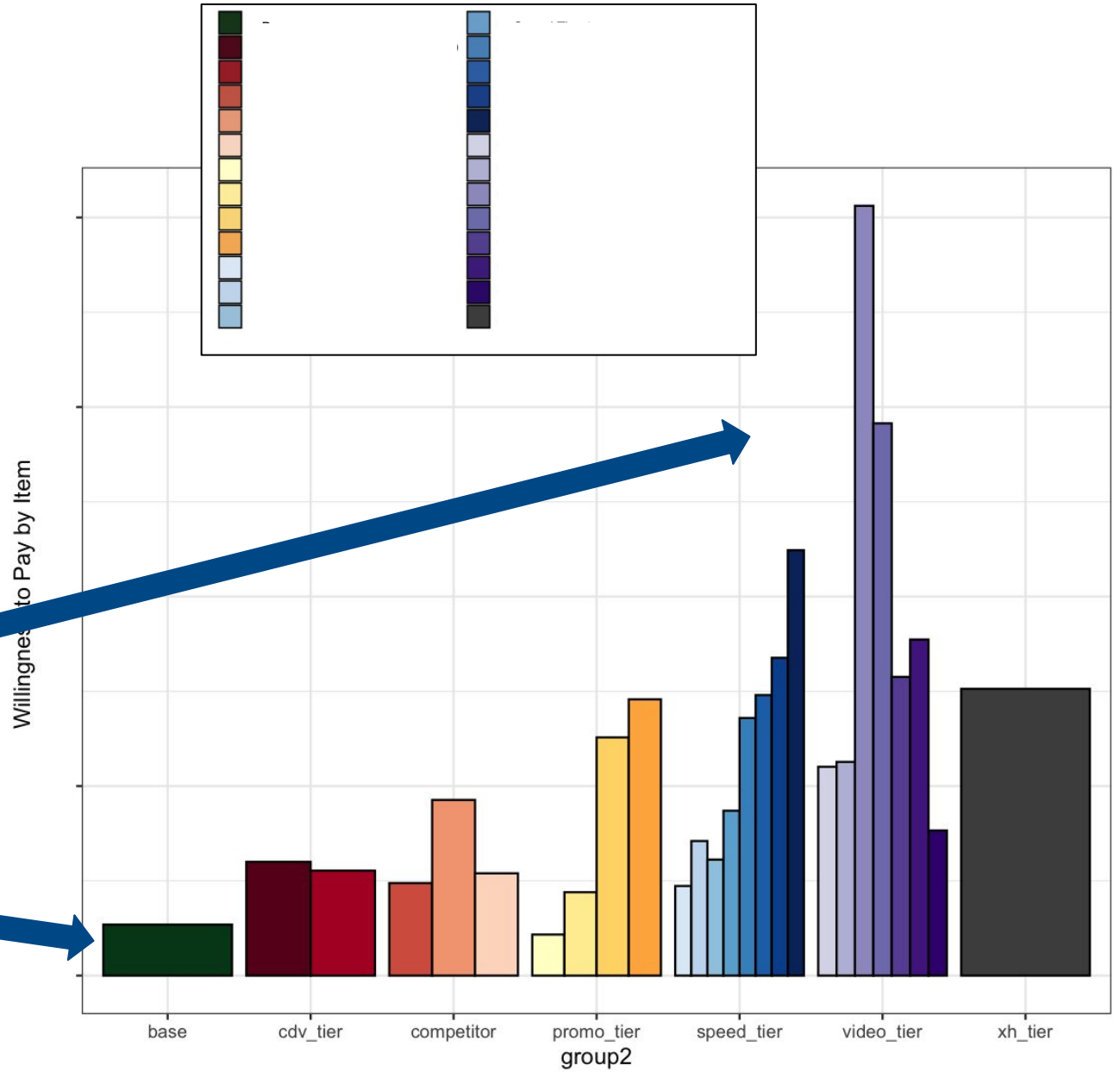


Training the Regression by Region

Pricing Model Northeast Region

Differences from Base Model:

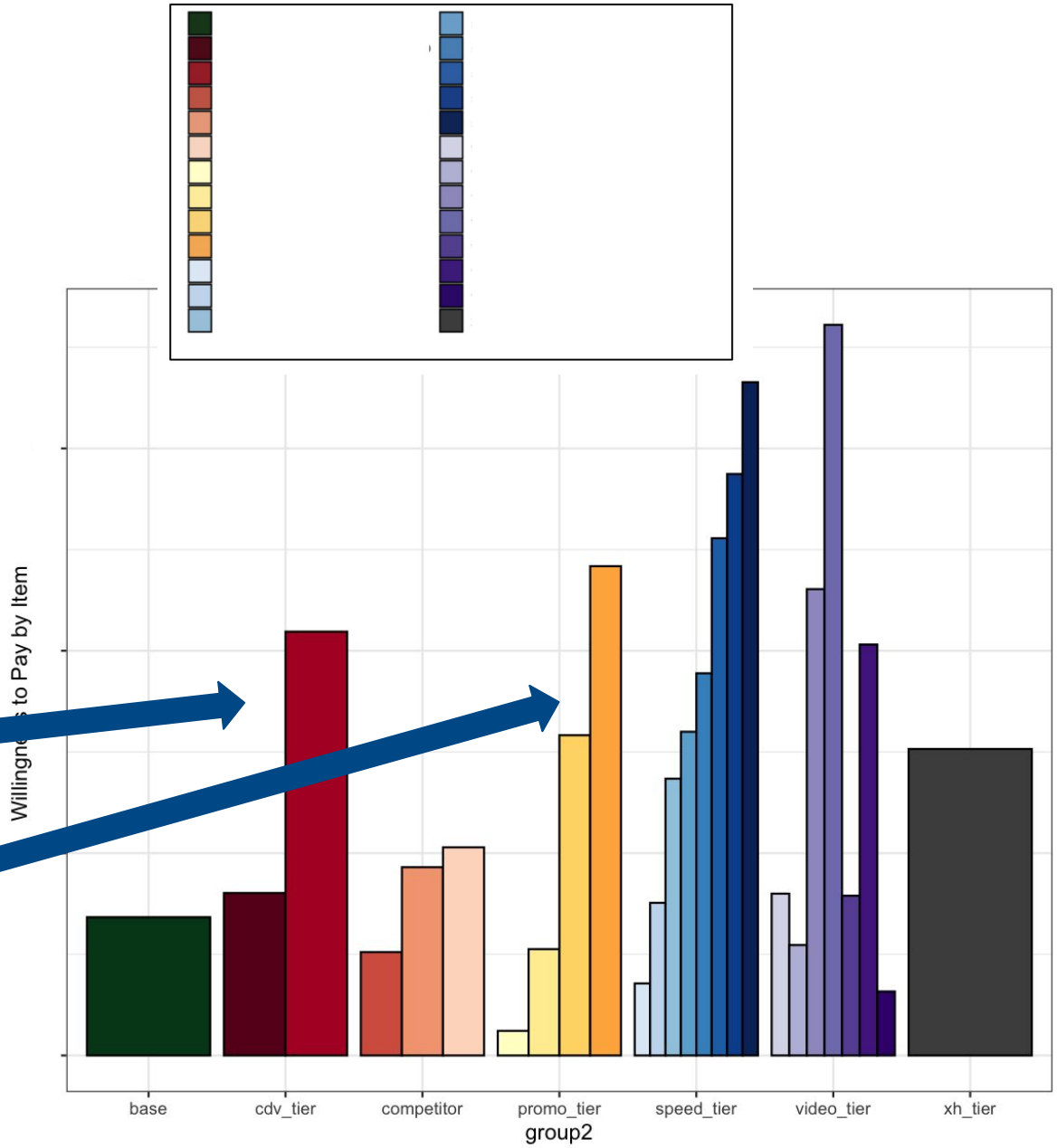
- Higher value for video tier 3 and 4
- Lower base price



Pricing Model Central Region

Differences from Base Model:

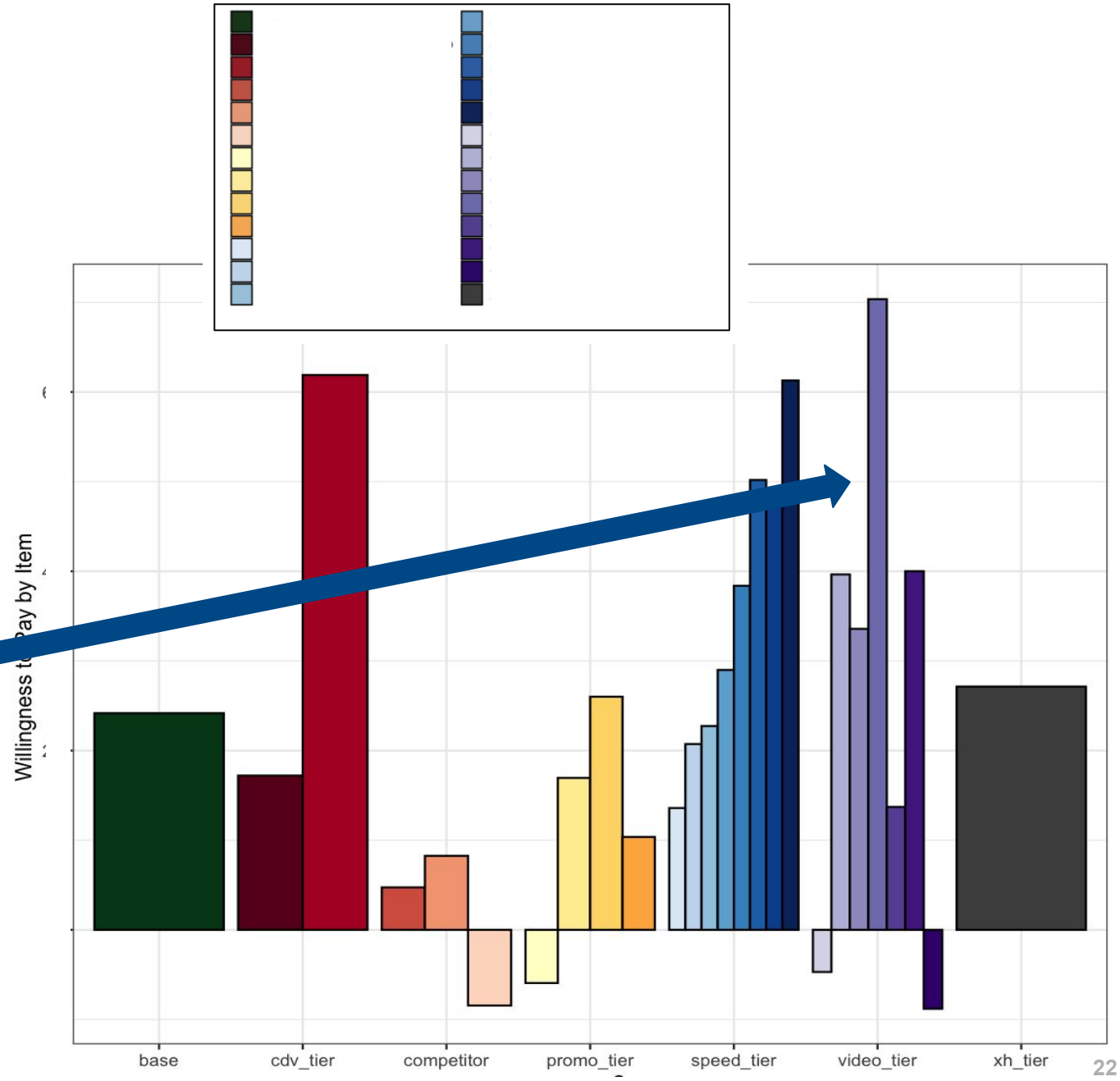
- Higher value in having CDV other
- Highest value on promo tier 3



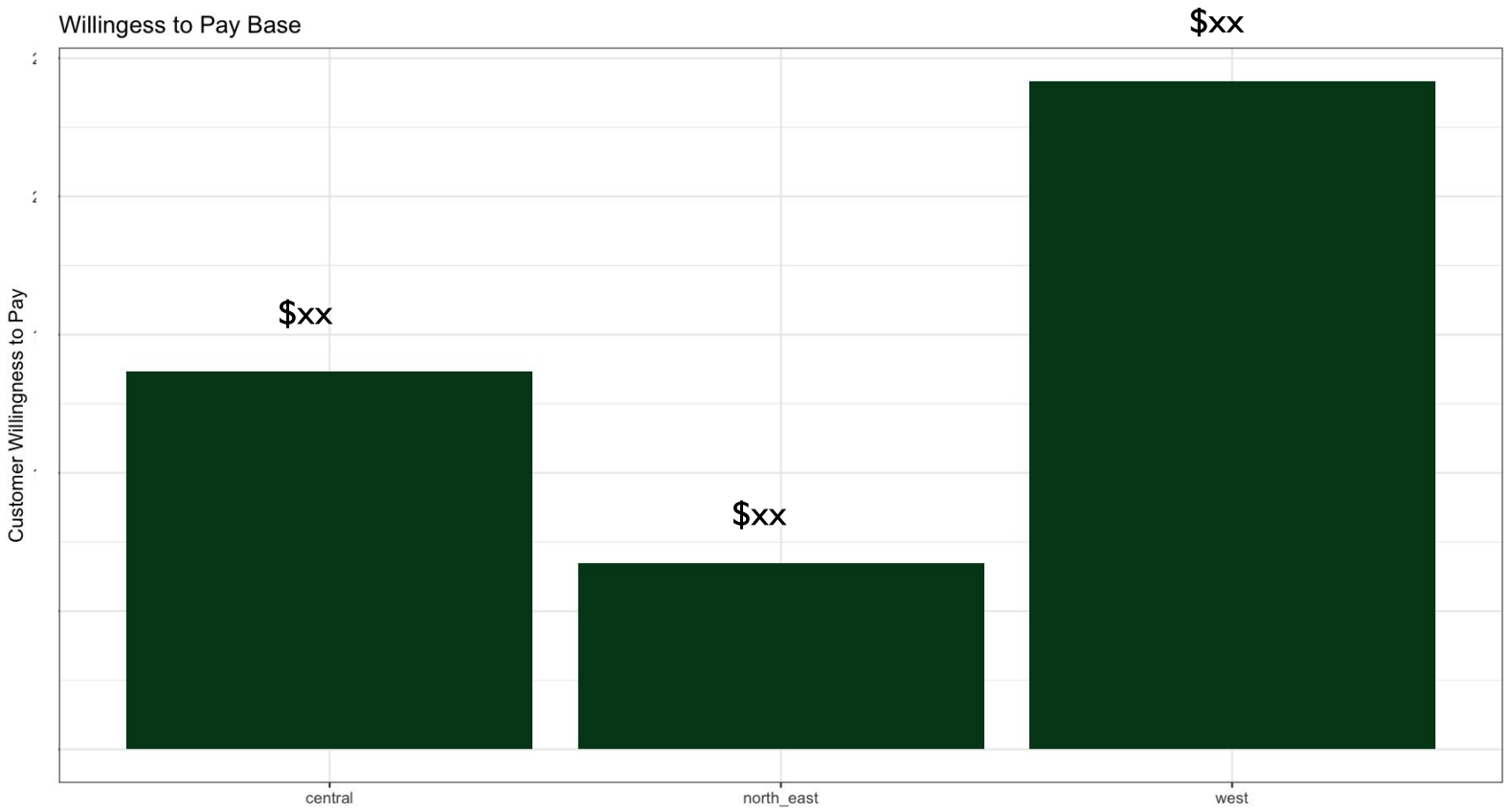
Pricing Model West Region

Differences from Base Model:

- Value video tiers 6-8 less than tier 4



The Base Price is Different in Each Region

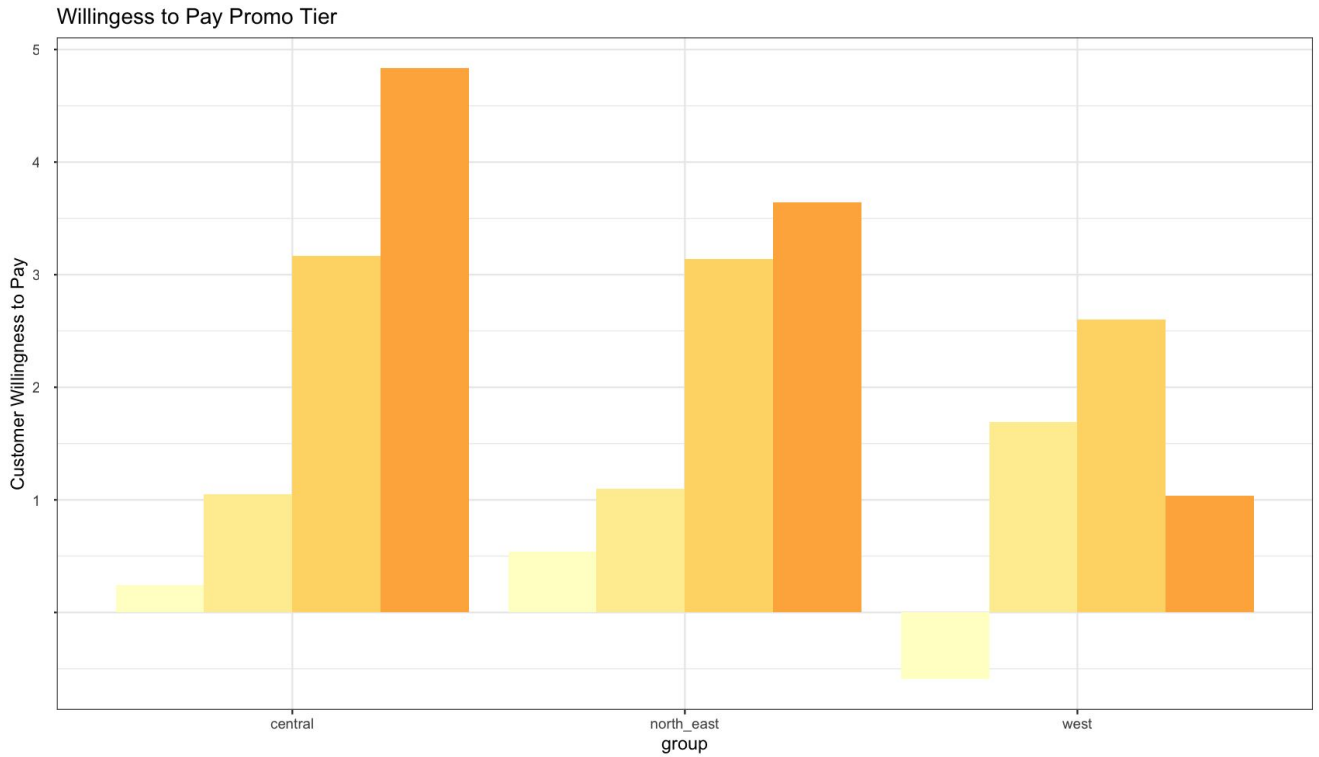


The base price indicates the price a customer is willing to pay before adding anything additional is added to the bundle. It assumes AT&T is the competitor

Base Package

1) No HSD
2) No CDV
3) No Video
4) Competitor AT&T
5) Promo Tier Blank

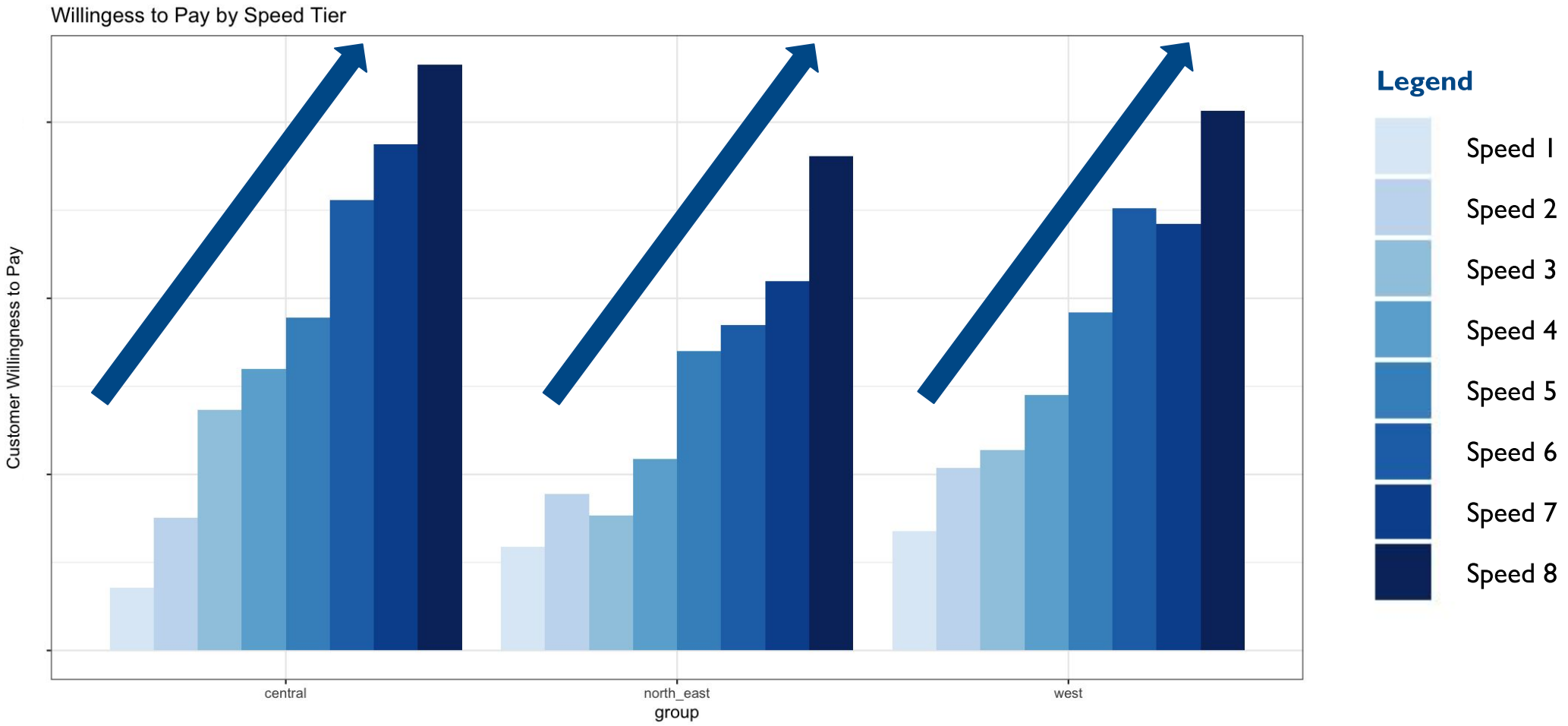
Increased Promotions has the Smallest Effect on West Region



Legend

- No Promotion
- Promo 1
- Promo 2
- Promo 3

Price Goes Up as Speed Tier Goes Up



Key Findings and Takeaways

Methodology | Clustering and Random Forest | Pricing Model | Findings and Takeaways

Customer Segmentation Recommendations

Further Segmentation

Video Tier & Speed Tiers are drivers of revenue

Segment customers by Video Tier & Speed Tier

Potential Upselling Opportunities

Video/Internet & Traditionalists with t3 tend to pay less than customers with t2

Digital tiers were the most popular amongst all clusters
(Traditionalists & All in One customers split between low and high digital)

Push customers to upgrade from t3 to t4

Push Traditionalists & All in One customers to upgrade to premium digital
(digital preferred plus & digital preferred video)

Key Findings from Supervised Learning Techniques

- Features with highest customer value:
  - Feature 1 (25% of bundle)
  - Feature 2 (19% of bundle)
  - Feature 3 (18% of bundle)
  - Feature 4 (15% of bundle)

- Features customers value to save money:
  - Feature 5 (24% of savings)
  - Feature 6 (17% of savings)
  - Feature 7 (13% of savings)
  - Feature 8 (12% of savings)
  - Feature 9 (9% of savings)

- Features with no/little bundle impact:
  - xx
  - xx
  - xx
  - xx
  - xx
  - xx
  - xx
  - xx
  - xx

Takeaways

1. Pricing remains fairly consistent across regions, but base WTP is highest in the West and lowest in Central

2. AT&T is the strongest competitor in the Central and West regions, while Verizon is more prevalent in the North East

3. Central Division values Video services much less than the West and Northeast

4. The Northeast has higher WTP for all internet. Central values ANY internet speed equally at lower tiers and values internet less overall.