CARMAX ANALYTICS SHOWCASE FINDINGS: PRICE, MODEL YEAR, MILEAGE

Charlie Frumkin, Sean Won, Esther Goldberg

DATA CLEANING

To apply quantitative analysis to numeric variables represented as categorical intervals, we converted the following attributes from ranges to single numeric values: price, appraisal offer, mileage, and appraisal mileage. For the sake of analysis, we replaced each interval with its midpoint (i.e., "$0 to $5k" → 2500), and we applied a similar conversion to data at the upper limits of these variables (i.e., "$40k+" → 42500). Despite some loss in accuracy, the midpoint conversion was the best way to apply continuous models to categorically represented data.

VARIABLE SELECTION

We performed most of our analysis in Weka, a software tool for machine learning. In Weka, we selected variables with the greatest explanatory power for price, model year, and mileage, the three purchased car attributes we deemed most useful for narrowing down customer recommendations. For each of these attributes, we used Pearson's correlation coefficient (Correlation Attribute Evaluation with Ranker method in Weka) to select the best appraisal attributes for our regression models. We then ran various regression classification algorithms using 10% cross validation, with the algorithm splitting the dataset into 10 chunks and rotationally using 9 of them to predict the remaining 1, then checking the degree to which the ten predicted chunks overlap.

ANALYSIS AND APPLICATION

While we obtained models that predict appraisal attributes from purchased vehicle attributes, we believe the most useful models are those that predict a purchased vehicle attribute from several appraisal attributes. Using such models, CarMax representatives will be able to use the characteristics of a customer's appraised vehicle to recommend vehicles in their inventory that match the price, model year, and mileage of vehicles the customer is predicted to purchase. This will allow for greater turnover of CarMax vehicles and thus more revenue for CarMax. Although we applied various machine learning algorithms, including linear models and decision trees, we found that linear regression models were most accurate and cost effective for making predictions based on the entire dataset. In our video presentation, we also discuss different types of models that were more accurate for subsets of the data, but linear models that best predicted the price, model year, and mileage of a purchased vehicle from the entire dataset are shown below.

1. Price = 0.2046*appraisal_offer + 243.658*model_year_appraisal – 516.196 *engine_appraisal + 30.746*horsepower_appraisal + 239.448*fuel_capacity_appraisal – 473563.825

(This regression model produced a mean absolute error of $6874.89 between predicted and actual prices, which falls within two $5k-intervals in the original dataset.)

2. Model year = 0.0794*model_year_appraisal + 0.0013*horsepower_appraisal + 1853.095

(This model predicted purchased vehicle model year with a mean absolute error of 1.5532 years.)

3. Mileage = 0.0127*appraisal_offer – 292.967*model_year_appraisal + 0.056* mileage_appraisal + 621196.1621

(This model predicted purchased vehicle mileage with a mean absolute error of 18807.27 miles.)