SEMMA Process Documentation for ML Case Study

BBY Retail Merchandise Chain

Prospect Customer Revenue prediction using ML

Ohm Kundurthy

2024

1. **Sample**

* **Libraries Imported**: Loaded essential R libraries (e.g., vtreat, dplyr, ggplot2, caret, gbm) for data manipulation, visualization, modeling, and evaluation.
* **Data Loading**: Read training, testing, and prospects datasets from CSV files, organized into multiple domains (consumer, donations, in-house, magazine, and political data).
* **Data Integration**: Merge datasets for training, testing, and prospects using left join on the key attribute, ensuring all relevant features are combined for each data partition.
* **Partitioning Note**: Skipped partitioning since the datasets were pre-sampled into training, testing, and prospects partitions.

1. **Explore:**

* **Examine Dataset:** Used summary() to explore possible values, data types for all attributes.
* **Visualization Summary:**

Table below summarizes different visualizations performed:

|  |  |  |
| --- | --- | --- |
| **S.No** | **Visualization Description** | **Inference** |
| 1 | Bar plot of ResidenceHHGenderDescription vs. average yHat. | Revealed highest predictions for "Cannot Determine" group; other groups showed similar averages. |
| 2 | Box plot of PresenceOfChildrenCode vs. yHat. | Showed limited predictive power with overlapping distributions. |
| 3 | Bar plot for ISPSA vs. yHat with correlation analysis (r = -0.0006). | Indicated a weak relationship. |
| 4 | Box plot for HomeOwnerRenter vs. yHat. | Showed no meaningful differences in distribution. |
| 5 | Scatter plot of NetWorth vs. yHat. | Revealed a general trend of decreasing predictions with increasing net worth. |
| 6 | Bar plot for Investor vs. yHat. | Highlighted no effect on predictions. |
| 7 | Box plot for BusinessOwner vs. yHat. | Showed no significant differentiation in distribution. |
| 8 | Bar plot of Education vs. yHat. | Identified a few standout categories with extreme mean values. |
| 9 | Box plot of HomeOffice vs. yHat. | Revealed similar distributions across groups. |
| 10 | Bar plot of stateFips vs. yHat. | Exhibited narrow distribution ranges across states. |
| 11 | Mapped yHat using latitude/longitude. | Highlighted spatial trends. |

* **High Cardinality Identification:** High-cardinality variables and irrelevant demographic data were identified for potential exclusion.
* **Occupation Industry & High Cardinality Analysis:** High cardinality identified in OccupationIndustry (19 levels), FirstName (4463 levels), and other variables
* Many demographic features such as EthnicDescription need vetting for use on ethical standpoint for machine learning.
* **Chi-Square Tests:** Statistical tests for categorical variables like PartiesDescription and ReligionsDescription showed no significant relationships with yHat.

1. **Modify**

* **Feature Standardization with vtreat:** Applied vtreat to transform and encode selected features while preserving predictive relationships.
* Weak or no significant relationships were found between several categorical variables and the target (yHat).
* Spatial trends (mapped via lat/lon) showed no actionable patterns.

1. **Model (Initial Fit)**

* **Dataset Preparation:** Defined the target variable (yHat) and predictor variables (predictors) from the training dataset.
* **Cross-Validation Setup:** Used 3-fold cross-validation (trainControl) to ensure robust model evaluation.
* **Evaluation Metrics:** Created a custom function (calculateMetrics) to compute performance metrics: RMSE, R², MAE, and MAPE.
* **Random Forest Model:** Trained a random forest model (rfModel) with 25 trees and 3 predictors per split. Retrieved cross-validation predictions, calculated performance metrics.
* **Gradient Boosting Model:** Built a gradient boosting model (gbmModel) with 25 trees, a learning rate of 0.1, and maximum depth of 3. Computed metrics using cross-validation predictions and actual values.
* **Linear Regression Model:** Trained a linear regression model (lmModel) and assessed performance via comparing actuals with cross-validation metrics.
* **XGBoost Model:** Developed an XGBoost model (xgbModel) with a depth of 3, learning rate (eta) of 0.1, and additional hyperparameter tuning. Evaluated predictions and computed metrics from cross-validation results.

1. **Assess**

* **Compare Metrics Across Models:** Metrics for all models (Random Forest, Gradient Boosting, Linear Regression, and XGBoost) were compared to establish a baseline understanding of error in predictions. Results in tabular format below:



* **Feature Importance**: Computed feature importance using varImp() for all 4 models
* **Consolidation of Important Features**: Gathered a comprehensive list of unique influential features across 4 models

1. **Modify (again)**

* **Feature Selection**: Retained only uniqueFeatures identified in previous step.
* **Feature Engineering**: Created new features:
  + **AnyDonation**: Identified if any donation was made across multiple categories (1 for any, 0 otherwise).
  + **HighSpenders**: Flagged households with upscale or collectible purchases.
  + **HasPet**: Consolidated pet ownership types into a single indicator.
  + **Categorical Re-Binning**: Grouped OccupationIndustry into broader categories for better predictive power.
  + **Continuous Feature Consolidation**: Summed numeric values extracted from magazine subscription data into a single continuous variable, MagazineSubscriptionCount.
* **Data Cleaning**: Ensured datasets were free of missing values before modeling.

1. **Model (Again)**

* **Data Preparation**: Target variable and predictors are defined. Missing values are handled using median imputation for both training and testing datasets.
* **Refit** Rebuild all 4 Models and gather metrics for both training and test datasets to evaluate predictive accuracy.

1. **(Refit) Assess**

* **Compare Metrics:** Metrics for all models (Random Forest, Gradient Boosting, Linear Regression, and XGBoost) were compared from training and testing data sets. Results are in tabular format below:

**A table with numbers and a black background

Description automatically generated**

1. **Predict on Prospect Data set**

* Identify predictors in the treated prospects dataset to ensure consistency with model requirements
* **Model Predictions**: Applied Gradient Boosting (gbmModelFinal), Linear Regression (lmModelFinal), and XGBoost (xgbModelFinal) models to generate individual predictions for each prospect.
* **Ensemble Prediction**: Combined predictions from the three models by calculating the row-wise mean (ensemblePred), leveraging ensemble learning for improved prediction accuracy.
* **Data Visualization:** Analyzed the distribution of ensemblePred using a density plot and histogram and compared it to the distribution of the target variable (yHat) in training data for validation.
* **Top Prospect Selection:** Ranked prospects by ensemblePred in descending order and selected the top 100 prospects for targeted outreach, saving the results to a CSV file for further use.

1. **EDA on Top Prospects**

* Table below is just a small portion of top 5 prospects by their revenue estimates.
* Many of these features exhibit the same or similar values across the top 100 prospects data which allows us to define the persona of a potential top prospect
* DogOwner\_lev\_x\_Yes: Most cases the value is 1
* EthnicDescription\_catN: Commonly has the value Other across many rows.
* MedianEducation Years: Values are consistently greater than 14
* PresenceOfChildrenCode Variables: Mostly 0, indicating absence of children.
* UpscaleBuyerInHome Variables: Commonly marked with 0 for both catN and catD

1. **Conclusion**

This study applied various machine learning models to analyze household and consumer data, focusing on prediction accuracy and feature importance. Exploratory data analysis (EDA) revealed that features like "DogOwner" and "MedianEducationYears" were strong predictors for identifying top prospects, while other variables, such as "OccupationIndustry" and political affiliation, showed weak correlations. Among the models tested (Random Forest, Gradient Boosting, Linear Regression, and XGBoost), Gradient Boosting proved most effective, with ensemble learning enhancing accuracy. Future research could explore the inclusion of more granular behavioral features and assess the models in real-world settings, while also addressing ethical concerns around sensitive demographic data to ensure fairness and regulatory compliance.

**DogOwner\_lev\_x\_Yes:**

* Most cases the value is 1

**EthnicDescription\_catN:**

* Commonly has the value Other across many rows.

**MedianEducationYears:**

* Values are consistently greater than 14

**PresenceOfChildrenCode Variables:**

* Mostly 0, indicating absence of children in the household.

**UpscaleBuyerInHome Variables:**

* Commonly marked with values like 0 for both catN and catD