## Scenario

<https://www.pexels.com/photo/person-showing-bedroom-interior-237371/>

A picture containing indoor, floor, bed, wall

Description automatically generated

You are the lead marketing data scientist for Bedding Bathing & Yonder (BBY). BBY is an American chain of domestic merchandise retail stores with an online presence. The chain primarily operates stores throughout the United States. BBY offers a membership and loyalty program designed to give customers easy access to great benefits and discounts through email, and physical direct mail coupons tailored to the household shopping habits. Loyalty customers have shared some basic information about their household and as a result the company’s marketing department has excellent data assets for these customers.

Senior management has asked you ***to model household revenue*** from a predefined household data set of these loyalty customers. Since pleasing these customers and maximizing their revenue potential is important to the company’s success, the marketing team has also purchased additional data sets to enrich BBY’s internal household data.

In addition to the training data the marketing department has a test set of households used primarily used for marketing campaign evaluation. Once your predictive model is built, you will evaluate the model within the training set *and* upon the test data. Your business colleagues will expect model metrics for both data partitions.

Assuming the model is both accurate and consistent for both partitions, the marketing department has segmented a third set of prospects. The model needs to predict household spend among these prospects so that the campaign designers can tailor the ads according to the amounts. For example, a household that is expected to spend $20 may receive an offer of buy one item, get the second one half off to encourage higher spending. Similarly, a prospect household with a higher predicted spend, such as $250, may receive a coupon for 10% off to trigger the purchase behavior for a large single item.

Lastly, with careful EDA, you are also encouraged to share any data, or model insights that help shape the company’s understanding of its best customers.

**You are asked to examine 15000 households’ attributes and their corresponding BBY spending habits. After examination, you will need to build a predictive model(s) on the training set. Next, evaluate the model(s) for any insights (important features) and model accuracy (RMSE, r-squared etc) within the training set. Next, apply the model to a test set of households to ensure consistency. Once satisfied, predict household spend on the prospect’s household data. Next, create a PowerPoint for your data science peers with data information and insights, modeling methods employed, model results, and provide the predicted prospect values in a CSV. You are NOT expected to create marketing campaign suggestions based on the model(s)’ output because that is customarily created alongside other business departments. You will narrate and record your presentation which is not meant to be more than 10minutes in length. You will also create a business report focused on your analytical efforts for wider business executive consumption within the company (additional guidance below).**

## Data

Source: The data has been synthesized using existing proprietary household data sets from various third parties obtained for a single US community. Some variables have been randomized, others have been anonymized, and further de-identification is performed on the geolocation attributes (lat/lon these have been completely manufactured). Thus, the data is likely not able to be reconstructed in its initial proprietary form but is still representative in some regards. You are not graded solely on model accuracy since the data itself has been purposefully adjusted.

## Data Tables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| File Name | Rows | Columns | Description | Partition |
| consumerData\_training15K\_studentVersion.csv | 15000 | 26 | Consumer purchasing habits | training |
| DonationsData\_training15K\_studentVersion.csv | 15000 | 15 | Household donation history | training |
| inHouseData\_training15K\_studentVersion.csv | 15000 | 20 | Data BBY has as part of membership | training |
| magazineData\_training15K\_studentVersion.csv | 15000 | 10 | Household magazine subscription history | training |
| politicalData\_training15K\_studentVersion.csv | 15000 | 13 | Household political leanings | Training |
| consumerData\_testing5K\_studentVersion.csv | 5000 | 26 | Consumer purchasing habits | Test |
| DonationsData\_testing5K\_studentVersion.csv | 5000 | 15 | Household donation history | Test |
| inHouseData\_testing5K\_studentVersion.csv | 5000 | 20 | Data BBY has as part of membership | Test |
| magazineData\_testing5K\_studentVersion.csv | 5000 | 10 | Household magazine subscription history | Test |
| politicalData\_testing5K\_studentVersion.csv | 5000 | 13 | Household political leanings | Test |
| consumerData\_prospects6K\_studentVersion.csv | 5992 | 26 | Consumer purchasing habits | Prospect |
| DonationsData\_prospects6K\_studentVersion.csv | 5992 | 15 | Household donation history | Prospect |
| inHouseData\_prospects6K\_studentVersion.csv | 5992 | 20 | Data BBY has as part of membership | Prospect |
| magazineData\_prospects6K\_studentVersion.csv | 5992 | 10 | Household magazine subscription history | Prospect |
| politicalData\_prospects6K\_studentVersion.csv | 5992 | 13 | Household political leanings | Prospect |

All data tables have a unique identifier, "tmpID" which can be used to join the data. If joined properly, the training data will have 80 variables described below. It may not be the case all variables are useful or even ethical to use in a model.

## Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Example** | **Definition** |
| tmpID | 5126 | Unique Household Identifier |
| ResidenceHHGenderDescription | Female Only Household | Head of household gender |
| EthnicDescription | Chinese | ethnic description |
| BroadEthnicGroupings | German | secondary ethnic description |
| PresenceOfChildrenCode | Known Data | If known, children purhcases indicate a child in home |
| ISPSA | 8 | Index of Social Position for Small Areas (ISPSA) Index of Social Position for Small Areas (ISPSA) an exclusive lifestyle classification system providing household characterization by social class affiliation. ISPSA is a score weighted on several census factors related to education and occupation. |
| HomeOwnerRenter | Likely Homeowner | Home dwelling type |
| MosaicZ4 | Bohemian Groove | Persona Descriptions from Experian  - i.e. https://assets.cengage.com/gale/help/dnow/Mosaic/MosaicTypeK40\_DescPortrait.pdf |
| MedianEducationYears | 13 | Number of formal education in years |
| NetWorth | $50000-99999 | Binned, net worth of household |
| Investor | Yes | Indication of an investor in the household |
| BusinessOwner | Yes | Indication of a registered business owner in the household |
| Education | Grad Degree - Likely | Modeled highest formal education attained |
| OccupationIndustry | Unknown | Head of household professional industry |
| HorseOwner | Yes | Indication of a equestrian rider and/or owned horse |
| CatOwner | Yes | Indication of a cat(s) in household |
| DogOwner | Yes | Indication of a dog(s) in household |
| OtherPetOwner | Yes | Indication of a non cat/dog/horse pet in household |
| HomeOffice | Yes | Indication of a home work office |
| BookBuyerInHome | 3 book purchases in home | Household book purchases from book-scan |
| UpscaleBuyerInHome | 1 upscale merchandise purchase | Household luxury goods pruchase indicator |
| BuyerofAntiquesinHousehold |  | Household antique purchase indicator |
| BuyerofArtinHousehold | Yes | Household fine art purchase indicator |
| GeneralCollectorinHousehold | Yes | Household collector purchases indicator i.e. cards, stamps |
| BooksAudioReadinginHousehold | Yes | Audio book purchase indicator |
| ComputerOwnerInHome | Yes | Presence of a home computer |
| ReligiousContributorInHome | 1 religious contribution in home | Donation to a religious charity indicator |
| PoliticalContributerInHome | 1 political contribution in home | Donation to a political nonprofit, candidate or elected official |
| DonatesEnvironmentCauseInHome | Unknown | Donation to environmentally focused non-profit |
| DonatesToCharityInHome | Unknown | Donation to a non-profit of any kind in household |
| DonatestoAnimalWelfare | Yes | Donation to ethical treatment of animals nonprofit indicator |
| DonatestoArtsandCulture | Yes | Donation to arts, performing, and culture related nonprofit organizations |
| DonatestoChildrensCauses | Yes | Donation to child advocacy and welfare non profit organizations |
| DonatestoHealthcare | Yes | Donation to healthcare and wellbeing organiations |
| DonatestoInternationalAidCauses | Yes | Donation to international, non-US causes i.e. relief agencies |
| DonatestoVeteransCauses | Yes | Donation to conflict and war veteran support organizations |
| DonatestoHealthcare1 | Yes | Donation to healthcare and wellbeing organiations |
| DonatestoInternationalAidCauses1 | Yes | Donation to international, non-US causes i.e. relief agencies |
| DonatestoWildlifePreservation | Yes | Donation to environmentally focused non-profit |
| DonatestoLocalCommunity | Yes | Donation to county, zip and community organizations |
| FirstName | Isiah | Fictitious first name for the head of the household |
| LastName | Bechtelar | Fictitious last name for the head of the household |
| Gender | M | BBY loyalty member gender |
| Age | 49.55844156 | BBY loyalty member age |
| TelephonesFullPhone | (539)-471-5789 | Fictitious telephone number (US) |
| lat | 48.29635 | latitude |
| lon | -95.54055 | longitude |
| county | Beltrami County | County (administrative division of a state, providing certain local governmental services) name |
| city | Grygla | City (administrative municipality) name |
| state | Minnesota | US state name |
| fips | 27007 | Federal Information Processing System (FIPS) Codes for States and Counties.  FIPS codes are numbers which uniquely identify geographic areas. |
| stateFips | 27 | FIPS state code |
| HomePurchasePrice | $243,000 | County register record for home price in last sale |
| LandValue | $11,000 | County register land deed assessed value |
| DwellingUnitSize | 1-Single Family Dwelling | Occupancy Type |
| storeVisitFrequency | 7 | Number of times loyalty member has been in store in last 12 months |
| PropertyType | Residential | Household property type |
| EstHomeValue | $187,500 | Estimated home value including land according to third party, non-county register model |
| yHat | 123.0180871 | DEPENDENT VARIABLE - The average household spend with BBY in USD. Some noise has been added to the amounts for further anonymization |
| FamilyMagazineInHome | 2 family-oriented magazine purchases | Magazine subscription related to family relations in house |
| FemaleOrientedMagazineInHome | 1 female magazine purchase | Magazine subscription with female target in house |
| ReligiousMagazineInHome | 1 religious magazine purchase | Magazine subscription on religious topic in house |
| GardeningMagazineInHome | 2 gardening magazine purchases | Magazine subscription covering gardening, and landscaping in house |
| CulinaryInterestMagazineInHome | 1 culinary magazine purchase | Magazine subscription focused on fine dining, diet, cooking or baking |
| HealthFitnessMagazineInHome | 4 health and fitness magazine purchases | Magazine subscription focused on health and wellbeing, working out, yoga or other fitness/active/sport related activities |
| DoItYourselfMagazineInHome | 3 Do-It-Yourself magazine purchases | Magazine subscription focused on home improvement, woodworking, or other self initiated projects |
| FinancialMagazineInHome | 1 financial magazine purchase | Magazine subscription focused on retirement, investment opportunities or retail equity |
| InterestinCurrentAffairsPoliticsInHousehold | Yes | Magazine subscription focused on newworthy events, local or global in nature |
| PartiesDescription | Republican | Registered political party affiliation |
| ReligionsDescription | Protestant | Likely religious denomination |
| LikelyUnionMember | Yes | Likely union roster appearance |
| GunOwner | Yes | Registered gun owner |
| Veteran | Unknown | Likely foreign war veteran or member of the armed service |
| supportsAffordableCareAct | Support | Political support for nationalized US healthcare |
| supportsGayMarriage | Support | Political support for LGBTQ+ rights |
| supportsGunControl | Oppose | Political support for gun ownership |
| supportsTaxesRaise | Oppose | Political support for raising taxes among wealthy households |
| overallsocialviews | Conservative | Overall classification of political leanings in the head of the household |
| DonatestoConservativeCauses | Yes | Donates to traditionally conservative cause nonprofit organizations |
| DonatestoLiberalCauses | Yes | Donates to traditionally liberal cause nonprofit organizations |

*Further data enrichment is possible using FIPS code but is not required for the case. The US Census publishes data using the FIPS code and some online data sets exists for community level attributes. However, the FIPs in this data set are incomplete and do not get to the block level. As a result, additional geospatial work is not required to complete the case. Exclusive to the case, for those that are interested, some starting links are shared below.*

<https://data.world/datasets/fips>

<https://www.huduser.gov/portal/datasets/usps_crosswalk.html>

## FIPS Code Explanation

A picture containing diagram

Description automatically generated

*Source:* [*https://www.policymap.com/2012/08/tips-on-fips-a-quick-guide-to-geographic-place-codes-part-iii/*](https://www.policymap.com/2012/08/tips-on-fips-a-quick-guide-to-geographic-place-codes-part-iii/)

## Criteria for Success

|  |
| --- |
| Organization – Was the presentation well organized? |
| Delivery – Was the content delivered clearly and persuasively with the audience in mind? |
| Code Documentation – Was the data mined to support the conclusion? |
| Written Supplemental – Is it grammatically acceptable, organized and error free.  -Is the data supported clearly and coincides with the data, and narration while being contextualized with external information? |
| Data Mining Process – Overall, as a complete portfolio of work, is the topic interesting, organized, researched, supported and delivered effectively? Was CRISP-DM, SEMMA, or a similar workflow followed to organize the work? |

**Delivery and Narration Guidance**

You are not allowed to use an ai avatar, or speech creation for narration. While this is certainly useful technology in many instances, the purpose of the business case presentation is to improve *your* presentation skills. In a business setting you will still be expected to articulate your findings and not send an avatar for this type of business meeting. As a result services like <https://elevenlabs.io/> or <https://www.heygen.com/> are not permitted. Your “boss” or “audience” in the case will not accept these technologies.

**Written Supplemental Guidance**

Submit a document to represent the entirety of your presentation including the data, process, findings, and implications in a business setting. Thus it’s a professional report, anything less than a professionally written and organized report will be considered sub-optimal. Amazon for example doesn’t use PowerPoint and instead uses “6 pagers” to make business recommendations, as such some organizations prefer written information over presentations. The use of external and verifiable sources is expected to add context and support any component of the paper. **The minimum is 2 pages maximum is 5**. **Double spaced and 12 point font.**

Helpful tips:

Markdown is not encouraged. It is to be a professional report similar to amazon's 6 pagers though does not have to be as long. As a business report consumed throughout manty departments, code and screenshots are discouraged/to be limited and instead describe your intentions/problem statement, data aspects, and results/findings/implications in an organized manner. Lastly, outlines and bullets alone will not earn you robust marks.

Overall guidance:

You are expected to submit R code, a voiceover narration of a live business presentation (can be a standalone video file, or inserted within the powerpoint) , a set of slides, and a written supplemental. If you submit documents with links to youtube videos or cloud drive files, you must ensure all links are accessible. Links set to private which are not viewable or downloabable will automatically result in 0 for that section of the rubric.