**DSND Starbucks Capstone**

Kristofer Bitney, 2019

*Abstract: Starbucks provided simulated data containing information from a series of promotional offers sent to members of their rewards program. I estimate the causal effect of receiving and viewing each promotional offer using a two-way fixed effects model. I then explore the extent to which member characteristics predict whether customers will view and complete each type of offer using regularized logistic regression with a natural cubic spline tensor product basis.*

**Data**

**Research Design and Data Summary**

The data tracks member transactions over a 30 day period, during which Starbucks members were issued one of ten promotional offers on days 1, 8, 15, 18, 22, and 25. At each of the seven assignment days, about 75% of the members were assigned to receive one of the ten possible offers and about 25% of the members were assigned to a no-offer control group. Chi-square tests suggests assignment to each offer was equally likely (Table 1).

|  |  |  |
| --- | --- | --- |
| *Table 1: Chi-Square Tests on Treatment Group Assignment* | | |
| Day | Chi-Square | P-Value |
| 1 | 7.304 | 0.605 |
| 8 | 3.254 | 0.953 |
| 15 | 5.037 | 0.831 |
| 18 | 4.06 | 0.907 |
| 22 | 4.617 | 0.866 |
| 25 | 8.624 | 0.473 |

Starbucks provided very limited information about the data and the nature of the research design underlying data collection. Data includes offer characteristics, member characteristics, and an event history that tracks offer receipt, offer views, and purchases.

Although I cannot be certain the offers were randomly assigned, F-tests of member characteristics on each of the seven offer assignment days suggest assignment was likely random (Table 2). With the exception of the proportion of members whose self-identified gender is “other”, sample proportions are very similar. With respect to offer assignment on days 18, 22, and 25, non-binary gender proportions differed between groups. I control for all of these member characteristics in all of my statistical models to account for observed differences, however minor.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Table 2: P-values from F Tests on Treatment Group Assignment by Member Characteristic* | | | | | |
| Day | Age | Membership Date | Income | Female | Gender Non-Binary |
| 1 | 0.477 | 0.908 | 0.715 | 0.505 | 0.751 |
| 8 | 0.422 | 0.330 | 0.290 | 0.484 | 0.895 |
| 15 | 0.987 | 0.663 | 0.329 | 0.208 | 0.856 |
| 18 | 0.568 | 0.974 | 0.995 | 0.529 | 0.000 |
| 22 | 0.477 | 0.836 | 0.861 | 0.182 | 0.000 |
| 25 | 0.374 | 0.746 | 0.672 | 0.352 | 0.000 |

Each of the 10 distinct promotions has different features. Offer types include buy-one-get-one-free (BOGO), discount, and informational. Offers vary in validity duration from three days to ten days. For BOGO and discount promotions, duration refers to the time period during which a customer can take advantage of the offer. Customers must make purchases to use BOGO and discount offers. Spending requirements and savings values for BOGO and discount offers vary. Informational offers do not have spending requirements and do not offer savings amounts. For informational offers, validity duration is defined as the assumed period of influence rather than the period during which a customer can take advantage of the offer. Offers were issued using different subsets of mediums, which can include web, email, mobile, and social. Table 3 describes the characteristics of the offers.

Member characteristics include membership enrollment date, age, gender, and income. The sample includes 17,000 members. Of those, 2,175 are missing data on all characteristics except enrollment date. Table 4 provides means and standard deviations of member characteristics by assignment group on the first day of the experiment period. Compared to the average American, Starbucks members in this sample are older, have higher incomes, and are more likely to be males.

The event history lists offer issue times, times at which offers were viewed, offer completion times, and purchase values at six-hour intervals over the 30 days of the experiment.

**Outcomes**

For causal inference, the outcome of interest is revenue. I define revenue as customer purchases net of promotion-related savings. I measure the predictive effects of member characteristics on the probability that members will view or complete the first offer to which they are assigned. Offer viewing and completion are treated as two distinct binary outcomes, for which I train two distinct models.

**Format**

For causal inference, I molded the data into a panel format. I aggregated the data to one-day intervals to reduce the number of time periods from 120 to 30. The aggregation facilitates use of statistical models that rely on big-N asymptotic theory for consistency.

For prediction, I restrict the timeframe in the analysis to the first seven days of the experiment period, during which members were exposed to at most one offer. I collapsed the data into a wide format.

I assume that once an offer is assigned to a member, the member is affected by that offer for the remainder of the thirty (or seven) day period. This approach captures the full residual effect of changes in customer behavior subsequent to completion of an offer.

**Missing values**

To estimate the average treatment effects of promotional offers on revenue, I do not drop any observations. While explore effect heterogeneity, I drop members missing characteristics. Each member with missing values is missing values for all of the member characteristics, and it is therefore not possible to impute missing values with multiple imputation under a Missing at Random (MAR) assumption. Likewise, mean imputation can reduce model variance, but in doing so it induces bias and is therefore not useful in an experimental setting. I instead handle missing values using list-wise deletion. Because treatment assignment is assumed to be random, list-wise deletion can produce unbiased parameter estimates.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Table 3: Offer characteristics* | | | | | |
| **Offer ID** | **Type** | **Reward** | **Minimum**  **Purchase** | **Validity**  **Duration** | **Marketing Channels** |
| 1 | Discount | $5 | $20 | 10 | web, email |
| 2 | Discount | $3 | $7 | 7 | web, email, mobile, social |
| 3 | Discount | $2 | $10 | 7 | web, email, mobile |
| 4 | Informational | $0 | $0 | 4 | web, email, mobile |
| 5 | BOGO | $10 | $10 | 5 | web, email, mobile, social |
| 6 | Informational | $0 | $0 | 3 | email, mobile, social |
| 7 | BOGO | $5 | $5 | 7 | web, email, mobile |
| 8 | BOGO | $10 | $10 | 7 | email, mobile, social |
| 9 | BOGO | $5 | $5 | 5 | web, email, mobile, social |
| 10 | Discount | $2 | $10 | 10 | web, email, mobile, social |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Table 4: Day 1 Member Characteristics* | | | | | | | | | | |
| Offer Group | Age | | Income | | Female | | Male | | Gender Non-Binary | |
|  | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| 1 | 55 | 18 | $65,134 | $21,984 | 0.42 | 0.49 | 0.56 | 0.50 | 0.01 | 0.11 |
| 2 | 54 | 17 | $65,715 | $21,191 | 0.42 | 0.49 | 0.57 | 0.50 | 0.02 | 0.13 |
| 3 | 54 | 17 | $64,676 | $21,492 | 0.40 | 0.49 | 0.59 | 0.49 | 0.01 | 0.11 |
| 4 | 55 | 18 | $65,225 | $21,359 | 0.41 | 0.49 | 0.58 | 0.49 | 0.02 | 0.13 |
| 5 | 54 | 18 | $65,616 | $21,499 | 0.41 | 0.49 | 0.58 | 0.49 | 0.01 | 0.10 |
| 6 | 54 | 18 | $66,012 | $21,405 | 0.43 | 0.50 | 0.56 | 0.50 | 0.01 | 0.10 |
| 7 | 54 | 17 | $64,343 | $21,811 | 0.42 | 0.49 | 0.56 | 0.50 | 0.02 | 0.14 |
| 8 | 54 | 17 | $66,282 | $22,308 | 0.44 | 0.50 | 0.55 | 0.50 | 0.01 | 0.12 |
| 9 | 55 | 18 | $65,314 | $21,555 | 0.41 | 0.49 | 0.57 | 0.49 | 0.02 | 0.12 |
| 10 | 54 | 17 | $65,569 | $21,624 | 0.39 | 0.49 | 0.59 | 0.49 | 0.02 | 0.13 |
| Control | 55 | 17 | $65,443 | $21,624 | 0.41 | 0.49 | 0.58 | 0.49 | 0.01 | 0.13 |

**Causal Effects of Promotions**

I measure the effects of viewing and receiving offers. We can assume offers are only effective when members actually view them. Still, from Starbucks’ perspective, the effects of offer receipt may be more important. Starbucks can send offers and even influence member viewing decisions, but Starbucks cannot ensure members will view them.

|  |  |  |  |
| --- | --- | --- | --- |
| *Table 5: Effects of Viewing Offers* | | | |
|  |  |  |  |
| Offer Group | β | SE | P |
| 1 | 1.52 | 0.37 | 0.00 |
| 2 | 3.49 | 0.18 | 0.00 |
| 3 | 3.28 | 0.31 | 0.00 |
| 4 | 0.81 | 0.20 | 0.00 |
| 5 | 2.19 | 0.16 | 0.00 |
| 6 | 1.30 | 0.12 | 0.00 |
| 7 | 2.14 | 0.25 | 0.00 |
| 8 | 2.63 | 0.18 | 0.00 |
| 9 | 2.75 | 0.21 | 0.00 |
| 10 | 4.32 | 0.18 | 0.00 |
| Age | 0.00 | 0.00 | 0.19 |
| Membership Date | -0.54 | 0.03 | 0.00 |
| Income | 0.06 | 0.00 | 0.00 |
| Female | 0.80 | 0.09 | 0.00 |
| Gender Non-Binary | 0.54 | 0.24 | 0.03 |
| Intercept | 22.72 | 1.53 | 0.00 |

**Causal Effects of Viewing Offers**

I use pooled Two-Stage Least Squares (2SLS) to estimate the effects of offers on customers who both receive the offers and view them. The 2SLS estimator, in this context, provides a method for estimating the effects of viewing offers solely through the covariance between offer receipt and offer viewing. In the first stage of 2SLS, I use offer receipt dummies to predict the probability that members will view each offer. In the second stage, I use the predicted probabilities as variables in a model that measures the causal effects of offer viewing on revenue.

Results are reported in Table 5. Robust standard errors are clustered on member. Note that membership enrollment date is expressed in years and member income is formatted in units of $1,000.

All of the offers increased revenue from customers who viewed them, though some were multiple times more effective than others. The results indicate discount offers with purchase requirements of $10 or less are the most effective promotions. Informational offers are the least effective. Results are interpreted as dollars of revenue increased per day. To measure the effects over a thirty day period, multiply the parameter estimate β by 30.

**Causal Effects of Receiving Offers**

I use a two-way fixed effects model to estimate the average treatment effect of offer receipt on revenue. The fixed effects estimator absorbs person and time effects—equivalent to controlling for dummy variables for each person and day—to account for all observable and unobservable characteristics that do not vary over both member and day. Member characteristics in this data are static over time, and are therefore dropped from the model. Independent variables were limited to promotion dummies. Robust standard errors are clustered on member.

I also report estimates from models without person or time effects, with person fixed effects only, with time effects only, and with covariates. If random assignment occurred, person-level fixed effects and covariates will not substantively affect parameter estimates. Time effects may still affect parameter estimates. Results are interpreted as effects per day, and are reported in Table 6.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Table 6: Effects of Receiving Offers* | | | | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Simple | | | Person Effects | | | Time Effects | | | Two-Way Effects | | | Covariates | | |
| Offer Group | β | SE | P | β | SE | P | β | SE | P | β | SE | P | β | SE | P |
| 1 | 1.50 | 0.10 | 0.00 | 1.70 | 0.17 | 0.00 | 0.80 | 0.11 | 0.00 | 0.71 | 0.19 | 0.00 | 1.66 | 0.11 | 0.00 |
| 2 | 3.06 | 0.14 | 0.00 | 3.53 | 0.19 | 0.00 | 2.34 | 0.14 | 0.00 | 2.53 | 0.20 | 0.00 | 3.45 | 0.16 | 0.00 |
| 3 | 2.18 | 0.13 | 0.00 | 2.62 | 0.19 | 0.00 | 1.47 | 0.13 | 0.00 | 1.63 | 0.20 | 0.00 | 2.46 | 0.15 | 0.00 |
| 4 | 1.24 | 0.09 | 0.00 | 1.76 | 0.15 | 0.00 | 0.53 | 0.10 | 0.00 | 0.78 | 0.17 | 0.00 | 1.37 | 0.11 | 0.00 |
| 5 | 2.12 | 0.13 | 0.00 | 2.58 | 0.19 | 0.00 | 1.40 | 0.13 | 0.00 | 1.58 | 0.20 | 0.00 | 2.31 | 0.15 | 0.00 |
| 6 | 1.43 | 0.09 | 0.00 | 1.85 | 0.14 | 0.00 | 0.72 | 0.10 | 0.00 | 0.87 | 0.15 | 0.00 | 1.58 | 0.10 | 0.00 |
| 7 | 1.75 | 0.11 | 0.00 | 2.16 | 0.17 | 0.00 | 1.04 | 0.11 | 0.00 | 1.17 | 0.18 | 0.00 | 1.98 | 0.13 | 0.00 |
| 8 | 2.33 | 0.13 | 0.00 | 2.66 | 0.18 | 0.00 | 1.63 | 0.14 | 0.00 | 1.68 | 0.19 | 0.00 | 2.59 | 0.15 | 0.00 |
| 9 | 2.48 | 0.17 | 0.00 | 2.94 | 0.25 | 0.00 | 1.77 | 0.17 | 0.00 | 1.95 | 0.26 | 0.00 | 2.80 | 0.19 | 0.00 |
| 10 | 3.77 | 0.15 | 0.00 | 4.11 | 0.20 | 0.00 | 3.07 | 0.15 | 0.00 | 3.13 | 0.21 | 0.00 | 4.17 | 0.17 | 0.00 |
| Age |  |  |  |  |  |  |  |  |  |  |  |  | 0.00 | 0.00 | 0.20 |
| Membership  Enrollment Date |  |  |  |  |  |  |  |  |  |  |  |  | -0.55 | 0.03 | 0.00 |
| Income |  |  |  |  |  |  |  |  |  |  |  |  | 0.06 | 0.00 | 0.00 |
| Female |  |  |  |  |  |  |  |  |  |  |  |  | 0.81 | 0.09 | 0.00 |
| Gender Non-Binary |  |  |  |  |  |  |  |  |  |  |  |  | 0.60 | 0.25 | 0.02 |
| Constant | 1.51 | 0.06 | 0.00 |  |  |  |  |  |  |  |  |  | 22.93 | 1.52 | 0.00 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Person Effects | No |  |  | Yes |  |  | No |  |  | Yes |  |  | No |  |  |
| Time Effects | No |  |  | No |  |  | Yes |  |  | Yes |  |  | No |  |  |

**Predictive Effects of Member Characteristics**

I train models that predict the likelihood that a member will view or complete an offer within seven days. I use the models to explore how member characteristics can be used to predict the likelihood that members will view and complete the offers they receive.

**Methodology**

I estimated label probabilities using a logistic regression model, with L2 regulation to shrink the model’s effective degrees of freedom and control overfitting. The offer viewing model included all ten offers used in the A/B tests. The completion model excluded the two informational offers and the members who received them. Both models controlled for age, membership enrollment date, income, female gender, and non-binary gender.

I scaled the continuous variables (income, member enrollment date, and age) using a min-max scaling rule. Min-max scaling adjusts the range of a variable to by expressing its values as a proportion of the original range centered at the range’s minimum. The rule can be written for each value x in the variable of interest.

I expanded the feature space of the three continuous variables using univariate natural cubic splines. Cubic splines increase the flexibility and locality of a function by dividing the domain of the function into multiple windows, and allowing estimation of separate cubic functions within each window. The windows are divided by “knots”. The cubic functions are constrained to preserve first and second degree continuity at the knots. Natural splines reduce variance by constraining the curves at the boundaries to be linear rather than cubic. In this analysis, the number of knots was selected separately for each of the three variables. Knots were spaced uniformly over each variable’s domain.

To further expand the feature space, I formed a tensor product basis—the full set of interactions between all of the features, including the natural spline functions. This is known as a “saturated model” in the econometrics literature (Angrist & Pischke, 2009). Tensor product bases increase hyperplane flexibility but can introduce spurious structure (Hastie et al., 2009). I use L2 regulation and cross-validation to control unnecessary structure from arbitrary sources.

Ideally, I would estimate logistic regression using a loss function that penalized integrated squared second derivatives of the linear component of the model function. Such a penalty directly penalizes function curvature, allowing for arbitrarily flexible and smooth curves without overfitting. I was unable to find an existing Python implementation of that loss function, and writing one is outside the scope of this project. I instead control overfitting by limiting the number of spline knots. I use an L2 penalty to de-emphasize pathological curvature and further limit flexibility.

I also consider models using tensor product bases with global cubic polynomials in lieu of splines. Global cubic polynomials offer less flexibility but more parsimony. Each univariate global polynomial requires three degrees of freedom, while each of the least flexible univariate splines I consider requires four degrees of freedom. After forming a tensor product basis, use of the least flexible splines requires an additional 42 degrees of freedom more than the global cubic polynomials. Due to the L2 penalty included in the model’s log likelihood loss function, the effective additional degrees of freedom is less than 42.

**Model Selection and Performance**

Using ten-fold cross-validation and exhaustive grid search, I selected the model with the lowest expected prediction error among those with support in the model space. Expected prediction error was measured with a zero-one loss function (i.e. accuracy).

Grid search estimates the mean cross validation score of every hyperparameter combination in a defined set. I estimated 3,900 models for each outcome. I tried sixty inverse regulation penalty values spaced uniformly within [0.005, 0.3]. That set implies a range of regulation penalty parameters in . The degrees of freedom used by each natural cubic spline expansion was selected from , implying 2-5 interior knots and two boundary knots for each variable. Sixty models for each outcome used global cubic polynomials instead of splines.

K-fold cross-validation estimates expected prediction error: the expectation of the test error of a model across training samples drawn from the population of interest. In other words, expected prediction error treats the training sample as a random variable.

We are generally more interested in the test error of the model that we trained on our sample. Test error treats the training sample as fixed. The most reliable way to estimate test error is to estimate predictions using a separate sample that did not influence model training or selection (Hastie et al., 2009). This is sometimes referred to as extra-sample test error. I split the data randomly into two data sets, with 90% of the data used to train and validate models and 10% of the data left to test performance after final models were selected and trained.

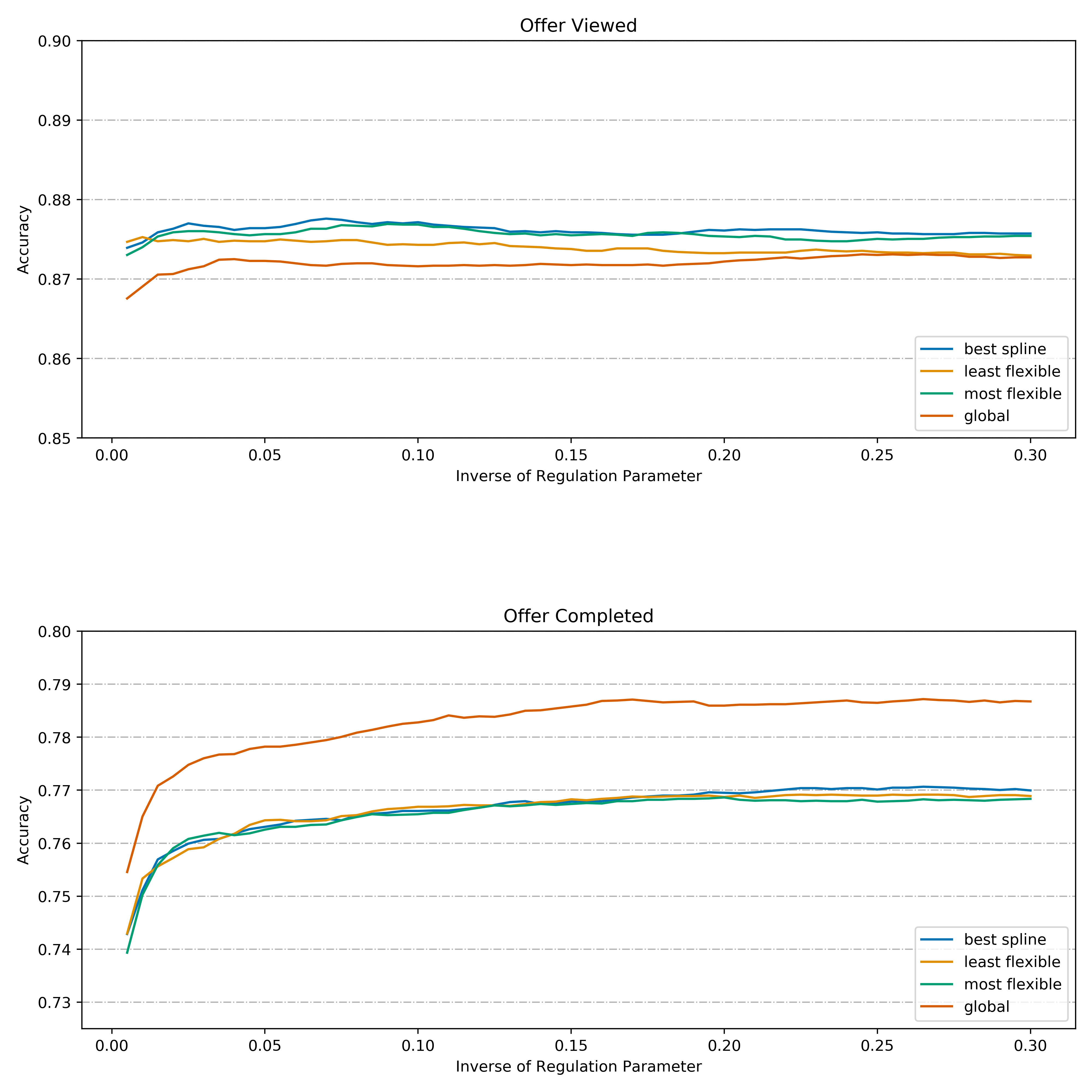
Figure 1 displays the mean cross-validation accuracy scores by inverse regulation parameter values for the best performing spline models, the least flexible spline models, the most flexible spline models, and global cubic polynomial models for both outcomes.

For the offer viewing outcome, the best performing model in cross-validation used a regulation penalty of 14.3 and splines for age, enrollment date, and income with 6, 4, and 6 degrees of freedom, respectively. The cross-validation accuracy score was 87.8%. Using the “one standard error” rule, I selected the most parsimonious model with mean cross-validation error that lies within one standard error of the best-performing model. This led to the least flexible spline model I considered, which used a regulation penalty of 200 and cubic splines each with three degrees of freedom. Model performance scores are listed in Table 7. The cross-validation accuracy of the selected model was 87.47%, the train accuracy was 87.7%, and the test accuracy was 87.5%.

For the offer completion outcome, the best performing model in cross-validation used a regulation penalty of 3.8 and global cubic polynomials. The cross-validation accuracy score was 78.7%. Using the “one standard error” rule led to a global polynomial model with a regulation penalty of 40. The selected model had a cross-validation accuracy of 77.5%, a train accuracy of 77.8%, and a test accuracy of 75.5%.

Test accuracy scores for the selected models by offer are listed in Table 8 for offer viewing and Table 9 for offer completion.

*Figure 1: Cross-Validation Performance: Mean CV Accuracy by Inverse Regulation Strength*



*Table 7: Model Accuracy in CV, Train, and Test Data*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Outcome** | **Model** | **Error Type** | **Sample Size** | **Accuracy** |
| Viewed | Best in CV | Cross-validation | 12,008 | 87.8% |
| Viewed | Selected | Cross-validation | 12,008 | 87.5% |
| Viewed | Selected | Train | 13,342 | 87.7% |
| Viewed | Selected | Test | 1,483 | 87.5% |
| Completed | Best in CV | Cross-validation | 10,230 | 78.7% |
| Completed | Selected | Cross-validation | 10,230 | 77.5% |
| Completed | Selected | Train | 11,367 | 77.8% |
| Completed | Selected | Test | 1,263 | 75.5% |

*Table 8: Test Accuracy by Offer ID for Offer Viewing*

|  |  |  |
| --- | --- | --- |
| **Offer ID** | **Sample Size** | **Accuracy** |
| 1 | 106 | 67.0% |
| 2 | 109 | 99.1% |
| 3 | 131 | 65.6% |
| 4 | 104 | 62.5% |
| 5 | 96 | 99.0% |
| 6 | 112 | 92.9% |
| 7 | 100 | 55.0% |
| 8 | 91 | 90.1% |
| 9 | 108 | 99.1% |
| 10 | 123 | 99.2% |

*Table 9: Test Accuracy by Offer ID for Offer Completion*

|  |  |  |
| --- | --- | --- |
| **Offer ID** | **Sample Size** | **Accuracy** |
| 1 | 107 | 64.5% |
| 2 | 97 | 71.1% |
| 3 | 120 | 62.5% |
| 5 | 106 | 60.4% |
| 7 | 118 | 61.0% |
| 8 | 124 | 60.5% |
| 9 | 112 | 72.3% |
| 10 | 109 | 71.6% |

**Results**

I estimated marginal effects of each variable and offer to explore the extent to which member characteristics predict the likelihood a member will view and complete each offer.

**Acknowledgements**

In addition to the material in Udacity’s Data Scientist Nanodegree courses, this paper relied heavily on Hastie et al. (2009). The analyses in this paper used the following Python libraries: Numpy, Pandas, Scikit-Learn, SciPy, Patsy, Statsmodels, Linearmodels, Matplotlib, Seaborn, and Jupyter.

**References**

Angrist, J.D. & Pischke, J.S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Hastie, T., Tibshirani, R., & Friedman, J.H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*.