FIFA 16 Purchase Prediction

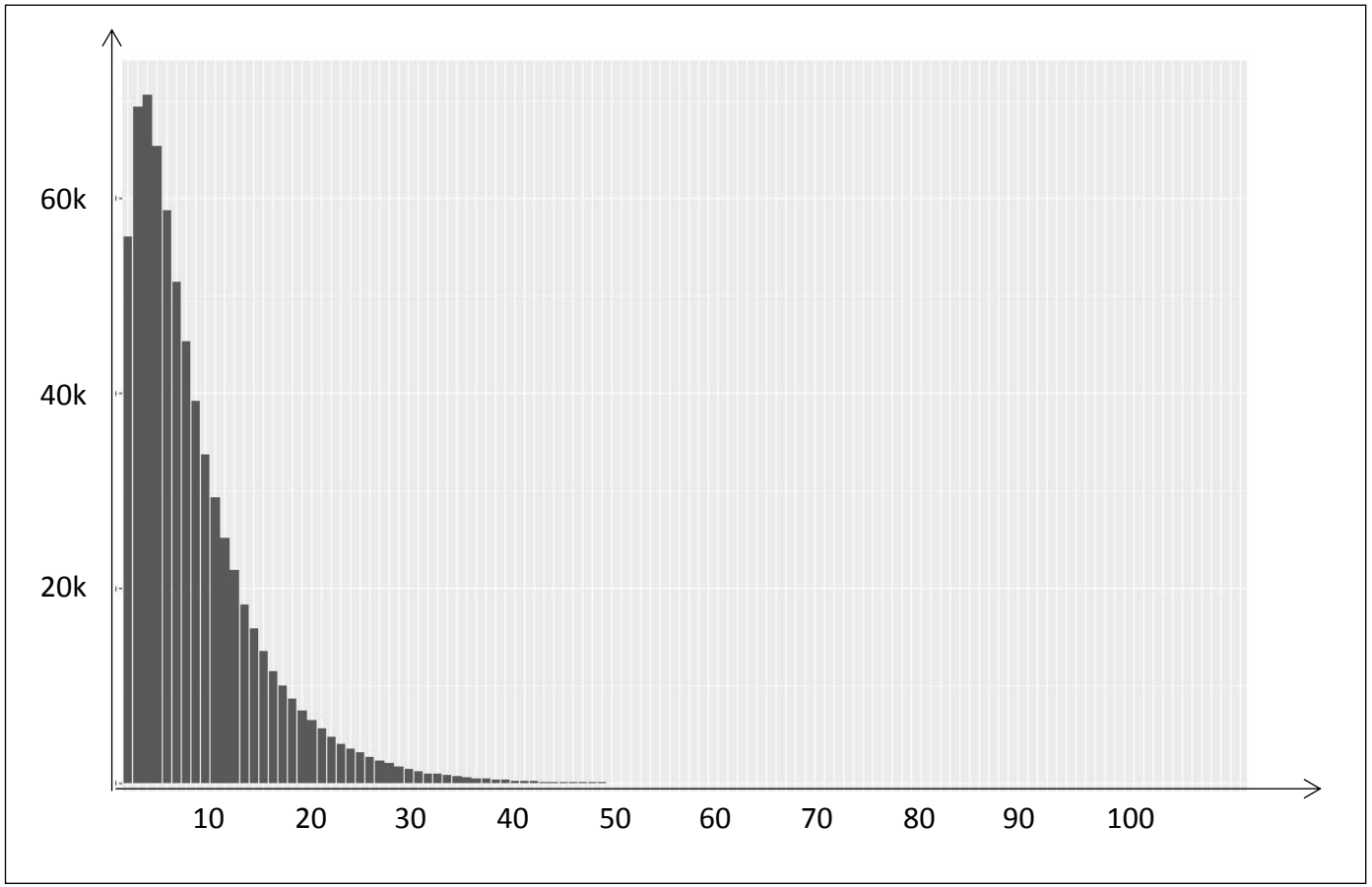
Sheng Zhang

Data Description

- The dataset contains 700,000 observations, 2 of which have NAs in every entry and were thus removed from the analysis.
- I also merged the dataset with a zipcode table to obtain geographical information (eg. state, city, longitude, latitude) for each player. Another 16 observations had zipcodes that were unidentifiable and were thus removed from the analysis as well.
- Therefore, the data that were used for the analysis had 699,982 observations, each representing a unique FIFA player.

Descriptive Data: User Characteristics

1 # of EA Titles Owned

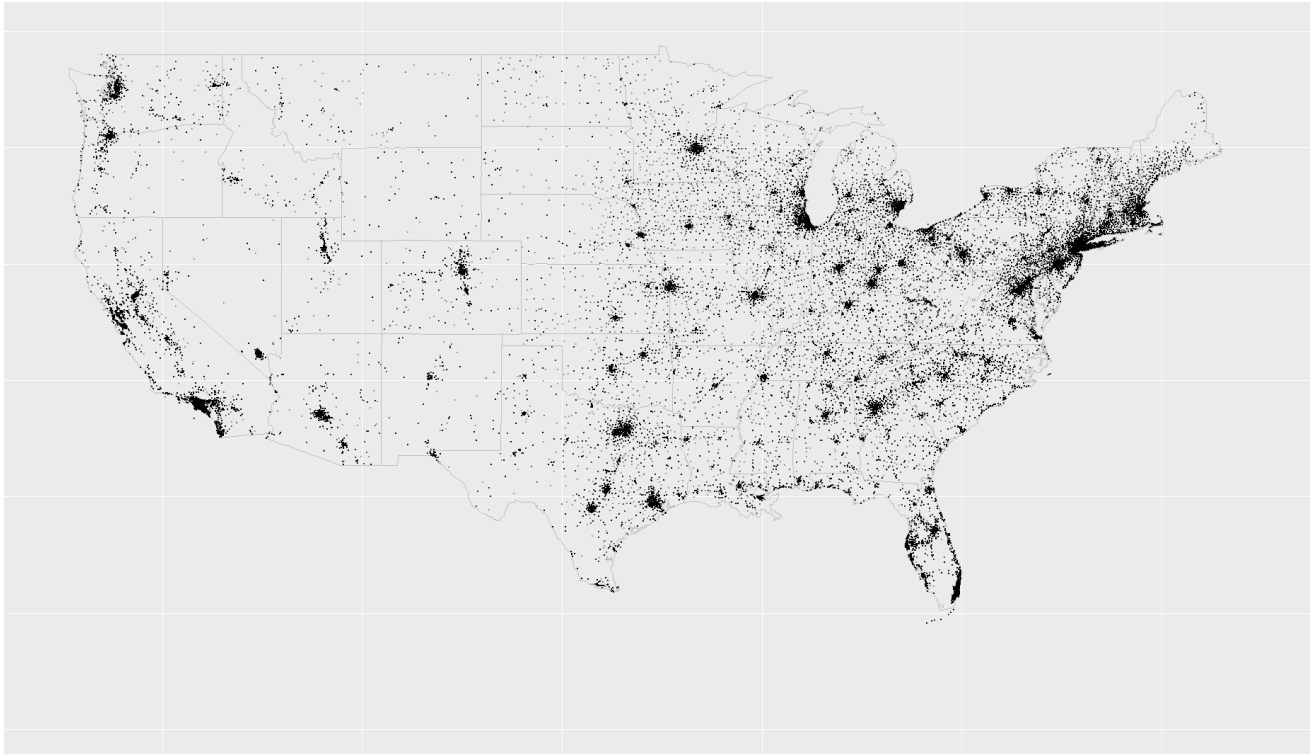


2 Platform Ownership

Platform	% Owned
PS3	46.84
PS4	18.07
X360	0
Xbox One	0
PC	9.67

Descriptive Data: Geographical Distribution

1 Map of Player Distribution

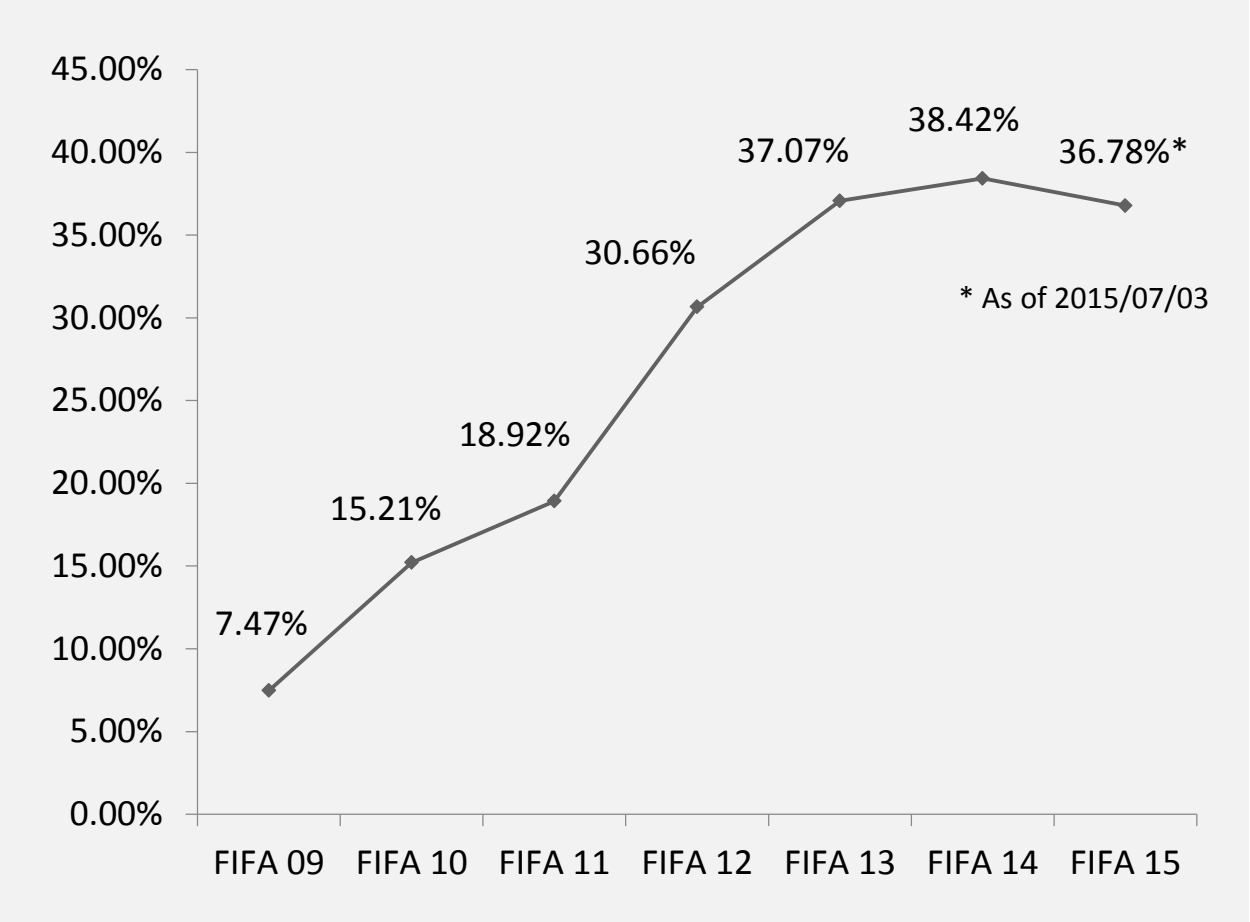


2 Top 10 States (by player count)

State	# of Players
CA	108,720
TX	61,803
NY	55,183
FL	42,286
NJ	35,904
IL	32,603
PA	29,209
VA	24,074
OH	23,170
MA	20,499

Descriptive Data: FIFA Popularity

1 % Purchased Each FIFA Version

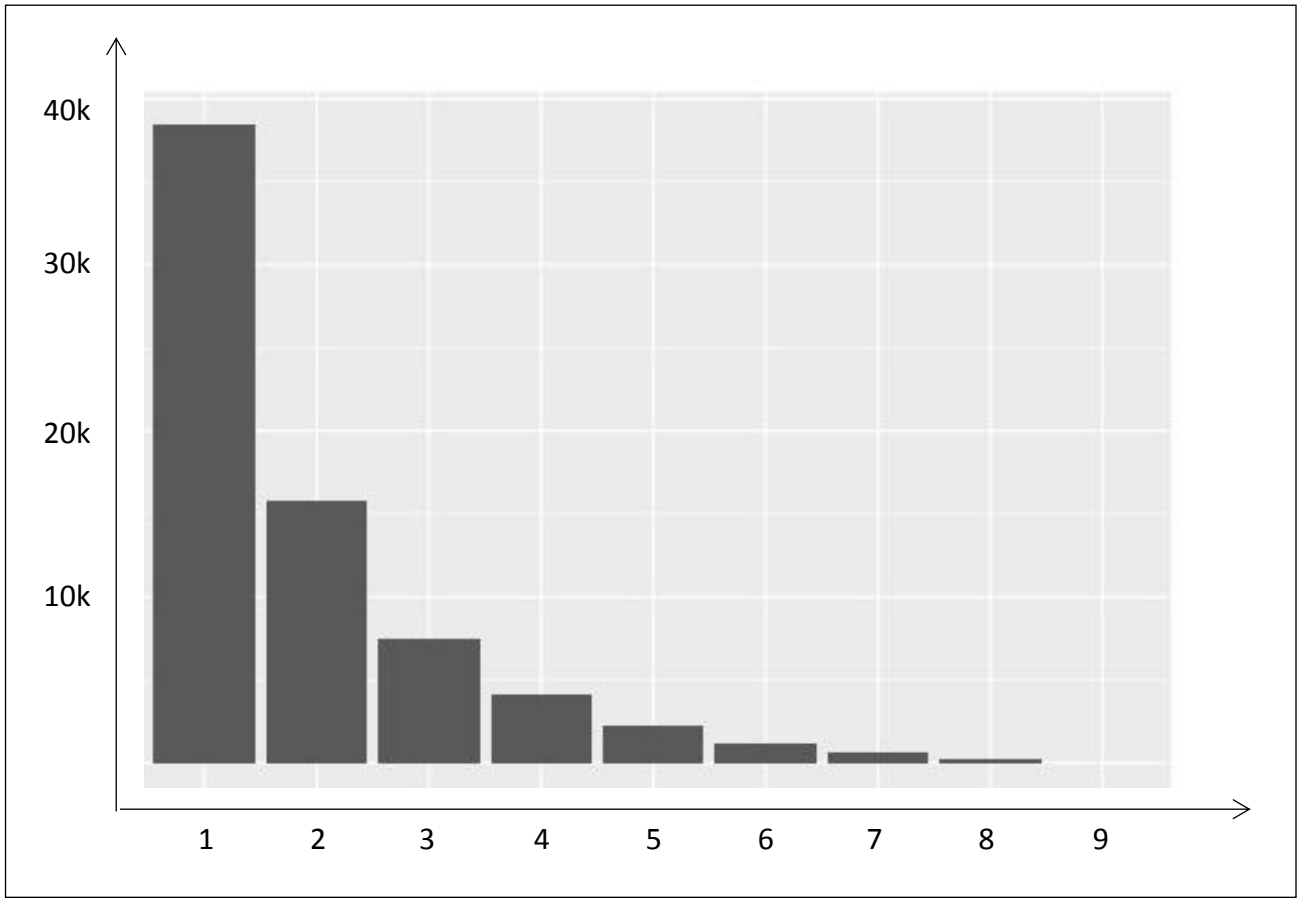


2 Top 10 States (by FIFA 15 purchase rate)

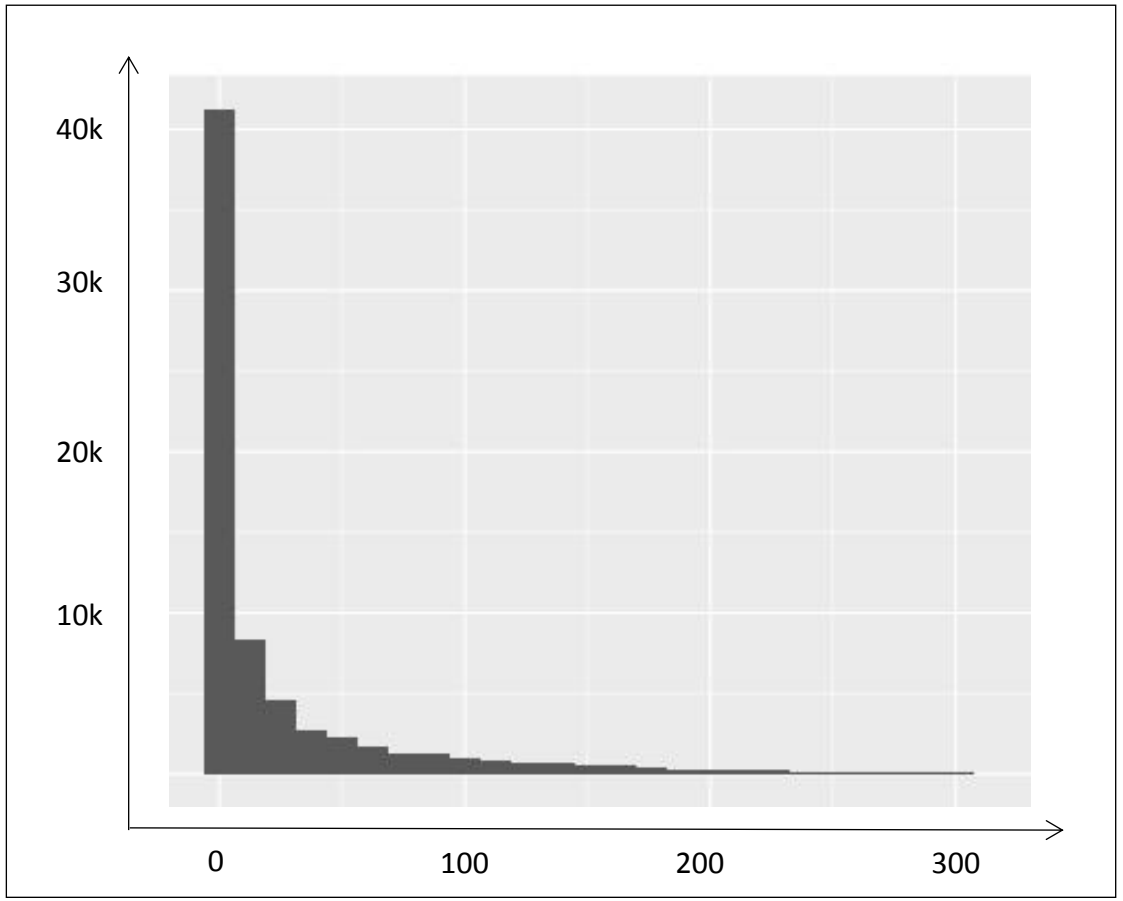
State	% Purchased FIFA 15
NV	44.40
VA	41.73
AZ	41.32
DC	40.09
FL	39.85
WY	39.37
WA	39.04
CO	38.84
CA	38.69
GA	38.39

Descriptive Data: FIFA Engagement

1 # of FIFA Titles Owned



2 Active Days in FIFA Franchise in Past 365 Days



Models: Features

- P_{i2015}: dummy variable on whether the user purchased FIFA 15, transformed from *F15* (date of purchase).
- P_{i2014}: dummy variable on whether the user purchased FIFA 14, transformed from *F14* (date of purchase).
- P_{i2013}: dummy variable on whether the user purchased FIFA 13, transformed from *F13* (date of purchase).
- P_{i2012}: dummy variable on whether the user purchased FIFA 12, transformed from *F12* (date of purchase).
- P_{i2011}: dummy variable on whether the user purchased FIFA 11, transformed from *F11* (date of purchase).
- P_{i2010}: dummy variable on whether the user purchased FIFA 10, transformed from *F10* (date of purchase).
- P_{i2009}: dummy variable on whether the user purchased FIFA 09, transformed from *F09* (date of purchase).
- $active\_rate_{it}$: the percentage of days that the user *i* is active in FIFA franchise during a "game year" *t*, transformed from *day_cnt*.
  - A "game year" is defined as the period between the release dates of two consecutive FIFA games. For example, the game year for FIFA 14 (t = 2014) is between 2013/09/09 and 2014/09/23.
- $ea\_titles\_over\_max_{it}$: the ratio of the user's EA titles owned compared to the maximum EA titles owned in the sample, transformed from *all_cnt*.
- $fifa\_titles\_over\_max_{it}$: the ratio of the user's FIFA titles owned compared to the maximum FIFA titles owned in the sample, transformed from *fifa_cnt*.
- $PS3_i$, $PS4_i$, $X360_i$, $XONE_i$, PC_i: dummy variables for platform ownership.
- $state_i$: indicator variable of the state that the user is from, transformed from *postal_code.*

Models: General Setup

- Intuition: I will predict next year's purchase behavior of a user based on his or her purchase behavior for the previous 6 FIFA games, *active_rate* last "game year", *ea_titles_over_max* and *fifa_titles_over_max* last "game year", the state that he or she is from, and the platform he or she owns.
- Mathematically, for each player *i* in game year *t*:

$$P_{it} = f(P_{it-1}, P_{it-2}, P_{it-3}, P_{it-4}, P_{it-5}, P_{it-6}, active\_rate_{it-1}, ea\_titles\_over\_max_{it-1}, fifa\_titles\_over\_max_{it-1}, state_i, PS3_i, PS4_i, X360_i, XONE_i, PC_i) + \varepsilon$$

- First, I will train the models using P_{i2015} as the outcome variable. I will also split the dataset into a training set (80% of all observations) and a test set (20% of all observations) to assess the predictive accuracy of each model.
- Mathematically, for each player *i*:

$$P_{i2015} = f(P_{i2014}, P_{i2013}, P_{i2012}, P_{i2011}, P_{i2010}, P_{i2009}, active\_rate_{i2014}, ea\_titles\_over\_max_{i2014}, fifa\_titles\_over\_max_{i2014}, state_i, PS3_i, PS4_i, X360_i, XONE_i, PC_i) + \varepsilon$$

- Then I will use the f estimated from using P_{i2015} as the outcome variable to predict P_{i2016}.
- Mathematically, for each player *i*:

$$P_{i2016} = f(P_{i2015}, P_{i2014}, P_{i2013}, P_{i2012}, P_{i2011}, P_{i2010}, active\_rate_{i2015}, ea\_titles\_over\_max_{i2015}, fifa\_titles\_over\_max_{i2015}, state_i, PS3_i, PS4_i, X360_i, XONE_i, PC_i) + \varepsilon$$

- I chose two models as potential candidates for estimating f, specifically, random forests and logistic regression (with lasso regularization).

Models: Assumptions

- Assumption I: The active rate of a player is roughly the same in the "game year" for FIFA 14 and in the "game year" for FIFA 15, both of which can be approximated by the active rate given in the dataset, which is from 2014/07/03 to 2015/07/03. In other words, I assume here that $active\_rate_{i2014} = active\_rate_{i2015}$.
  - Improvement with more data: with raw data such as the date of being active in FIFA franchise for each player from 2009 to 2015, I can calculate the active rate for each player during any time period, thus improving the accuracy of my input feature.
- Assumption II: The ratio of owned EA items of each user over the maximum owned EA items of the entire sample does not change much from the "game year" for FIFA 14 to the "game year" of FIFA 15. In other words, I assume that $ea\_titles\_over\_max_{i2014} = ea\_titles\_over\_max_{i2015}$.
  - Improvement with more data: by knowing how many EA items each user purchased during the "game years" of FIFA 15 and FIFA 14, I can calculate the number of EA items owned by each user at the start of the "game year" of FIFA 15 and of the "game year" of FIFA 14, thus improving the accuracy of my input features.

Models: Random Forests

- Random forests are an ensemble learning method based on decision trees.
- Formally, in decision trees, classification trees in particular, we divide the predictor space (set of possible values of $\boldsymbol{X}$) into M distinct regions $R_1, R_2, \ldots, R_M$, and for every new observation in a region R_m, we predict that the value of y for it will be the mean of y values for all training observations in R_m.
- Random forests, just as in bagging, estimate a number of decision trees on random samples drawn from the training dataset. However, random forests also impose an additional restriction on the number of predictors that can be considered as candidates at each split in the decision tree.
  - The rationale is that not limiting the number of predictors will likely result in the same set of predictors (strong predictors) being selected in all of the decision trees estimated, leading to highly correlated bagged trees. As a result of this high correlation, the model built would not reduce the variance by a lot and will not solve the over-fitting problem.

Models: Logistic Regression (Lasso)

- Logistic regression is a regression model that has a categorical dependent variable. Formally, the model is specified as:

$$log \frac{p(Y = 1|\boldsymbol{X})}{1 - p(Y = 1|\boldsymbol{X})} = \boldsymbol{X}'\boldsymbol{\beta}$$

  In other words,

$$p(Y = 1|\boldsymbol{X}) = \frac{\exp(\boldsymbol{X}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}'\boldsymbol{\beta})}$$

- Lasso (least absolute shrinkage and selection operator) regularization is a method used for feature selection. It can shrink the coefficient estimates toward zero, thus yielding a sparse model, which involves only a subset of independent variables. Specifically, the lasso coefficients minimize the quantity:

$$||\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- The tuning parameter λ will be selected using cross validation (fitting multiple models with different λ, then use the λ that results in the best performance).

Models: Comparison

- The main advantage of logistic regression is that the model is relatively easy to understand and interpret. Moreover, it is relatively fast to implement. However, logistic regression tends to have problems with non-linear classification boundaries.
- Compared to logistic regression, decision trees work better when there are more than one underlying decision boundaries and when the classes approximately lie in hyper-rectangular regions. However, decision trees are more prone to over-fitting, an issue that methods like random forests can help resolve.

Results: Comparison

- After training the models with the training set (where P_{i2015} was the outcome variable), both the random forests and the logistic regression (with lasso regularization) produced predicted probability of purchase based on the input features in the test set. A binary prediction $\hat{P}_{i2015}$ is then produced by comparing the predicted purchase probability to the mean of P_{i2015} in the training set.
- I then assessed the predictive accuracy of each model by assessing whether $\hat{P}_{i2015}$ matched the real P_{i2015} in the test set.
- Random forests yielded a predictive accuracy of 98.62% on predicting P_{i2015}.
- Logistic regression (lasso) yielded a predictive accuracy of 96.37% on predicting P_{i2015}.
- Both models seem to perform really well, considering that the mean of P_{i2015} is around 36.78%, which indicates that a naïve model that always predicts no-purchase could only yield a predictive accuracy of around (1 – 36.78%) = 63.22%.
- Since both models perform really well, I proceeded and used both of them to predict P_{i2016}.

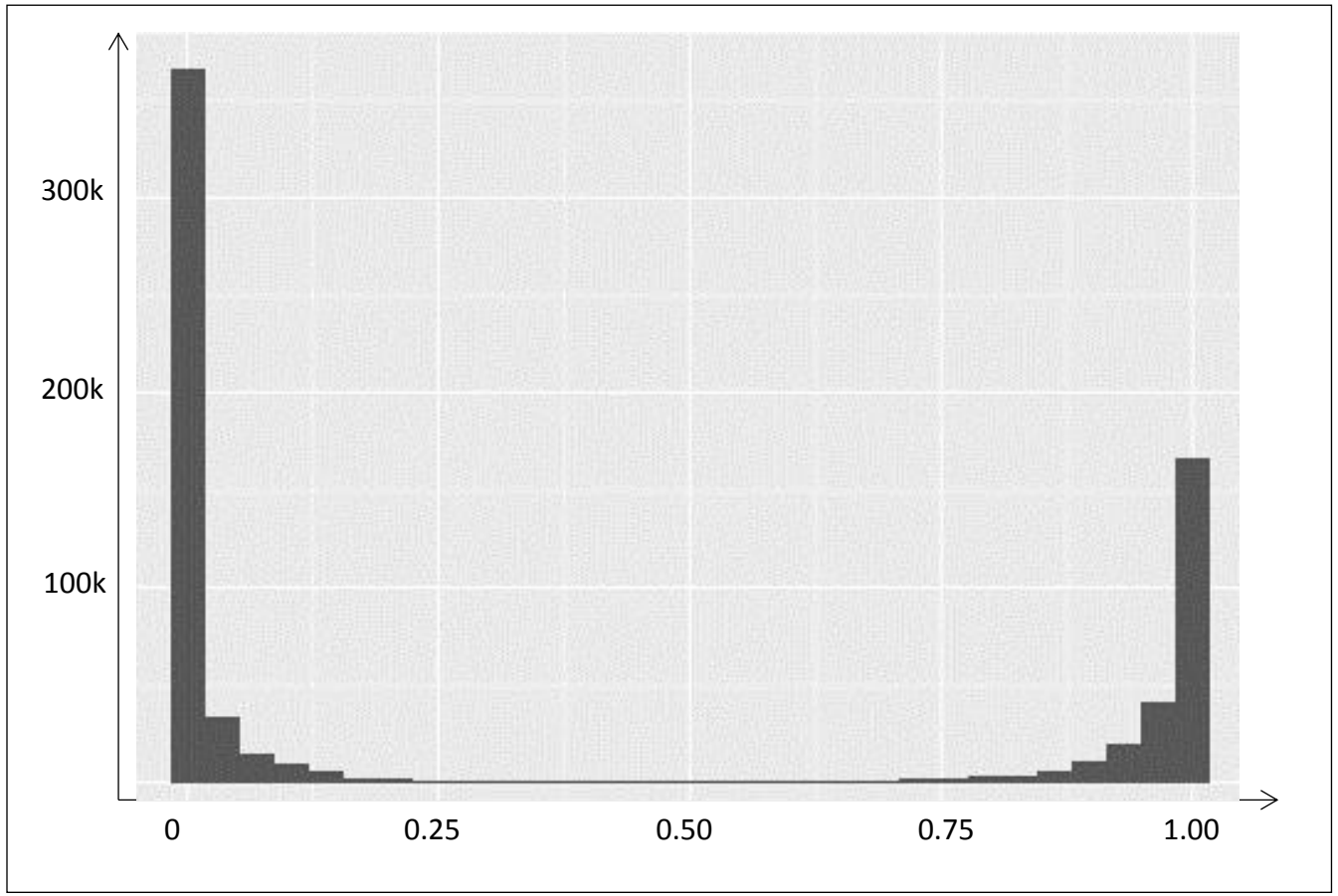
Results: Random Forests

1 Importance

Variable	% Increase MSE w/o Variable *
P_{it-1}	104.45
P_{it-2}	56.97
P_{it-3}	55.47
P_{it-4}	37.14
P_{it-5}	34.43
P_{it-6}	17.26
$PS3_i$	5.29
$PS4_i$	7.54
$X360_i$ & $XONE_i$	0
PC_i	17.38
$State_i$	2.28
$Active\_rate_{it}$	53.61
$EA\_titles\_over\_max_{it}$	5.83
$FIFA\_titles\_over\_max_{it}$	62.49

* Higher number indicates higher importance

2 Distribution of the Predicted Purchase Likelihood of FIFA 16

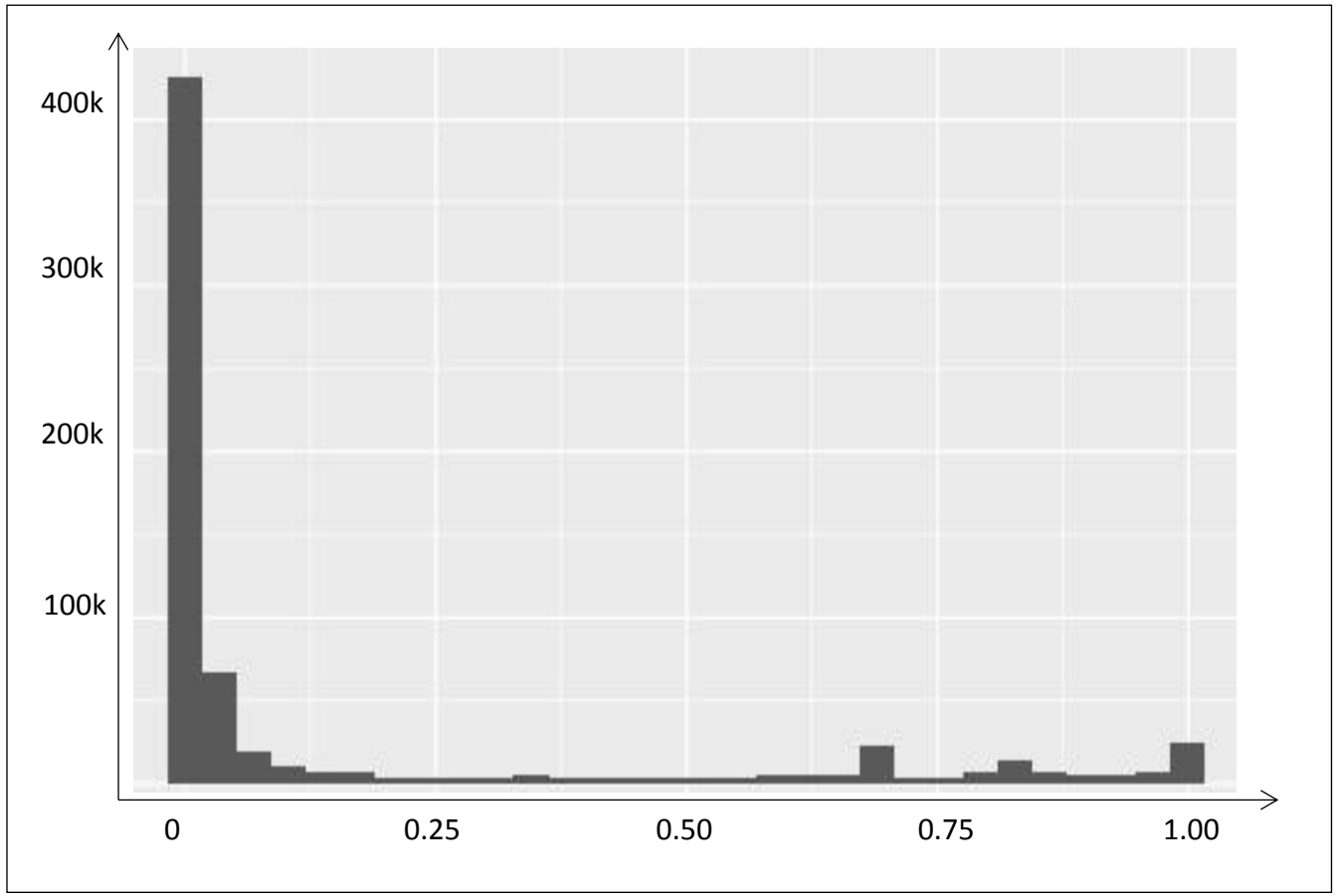


Results: Logistic Regression (Lasso)

1 Parameters

Variable	Coefficients
Intercept	-5.31
P_{it-1}	-6.46
P_{it-2}	-6.62
P_{it-3}	-6.95
P_{it-4}	-7.29
P_{it-5}	-7.92
P_{it-6}	-8.47
$PS3_i$	-0.86
$PS4_i$	1.14
$X360_i$ & $XONE_i$	0
PC_i	-0.38
$State_i$	0
$Active\_rate_{it}$	9.58
$EA\_titles\_over\_max_{it}$	-1.03
$FIFA\_titles\_over\_max_{it}$	62.92

2 Distribution of the Predicted Purchase Likelihood of FIFA 16



Discussion

- Future Steps:
  - The predicted results could potentially be more accurate if we had more data, as discussed in the assumptions section.
  - The coefficients of the logistic regression (lasso) mostly seem reasonable (+ for active rate, FIFA titles owned over max, and PS4; - for PS3 and PC.)
    - The reason why past purchases had negative coefficients is worth further investigation.
- Importance of geographical data:
  - Both algorithms indicate that the state indicator variable did not have much predictive power on future purchase behavior in my model, which included many other predictor variables. It is likely that the predictive power of geographical data is already picked up by other variables.
    - For example, it could be the case that players from some states own more EA titles.
  - Purchase rate by state (page 5) suggests that geographical variation in FIFA adoption does exist.
  - A standalone model with only geographical variable(s) as the predictor variable(s) could help determine whether geographical data by itself have any value in predicting sales. However, such a model will almost certainly perform worse (in terms of predictive accuracy) than the current models, which also used many other predictor variables.

Discussion

- Recommendation for media campaign targets:
  - Intuitively, we should focus more on the “likely” group (eg. Players with ~80% purchase probability)
    - Players who have very high purchase likelihood may purchase regardless of whether there are media campaigns, while players who have low purchase likelihood may not purchase regardless of whether there are media campaigns.
  - However, where we should draw this threshold depends on how much impact the media campaigns have, especially in terms of increasing purchase likelihood.
    - Results from past Marketing Mix Modelling (MMM) analysis could be useful here.
  - We also want to take into account of the different costs of reaching different kinds of customers to ascertain targeting which groups of customers leads to the largest return on investment (ROI).

Thank You!