GR5399: Statistical fieldwork

Internship Report

Student Name: Yunbo Zhu

Student UNI: yz3712

Internship start/end dates: 6/1/2020 – 8/21/2020

Total number of hours: 480

Internship sponsoring organization, department, and location:

Apple Inc.

Apple Media Product Data Science/Analytics – Internet Software & Services

Cupertino, CA

Supervisor Information:

Joe Li

Senior Data Scientist

+1 (408)996-1010

zli8@apple.com

Apple Inc. is an American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, computer software, and online services. It is considered one of the Big Tech technology companies, alongside Amazon, Google, Microsoft, and Facebook. I interned in division of Apple Media Product Data Science/Analytics under Internet Software and Services organization.

Among different line of businesses of Apple Media Products, I interned in video product team and my project title is TV app Engagement Patterns and Golden Behaviors. Not only streaming service applications, every product's goal is towards building something that has lasting value and entertainment, that is to say something you want to use every day. So, my motivation of this project is to let the TV app be the best TV screen for every moment and my goal is to understand how different in-app actions affect user engagement and how could we improve.

To achieve this goal, I start with understanding how customers use the application today. After identifying the most engaged users or to say our golden customers and how they use the application, we know where to focus to drive more of them and give recommendations to further improve the product.

With tons of data TV app collects every second, value is going to mean different things to different people: some folks will want to watch TV on their phone while commuting, some want to kick back on the couch and watch a movie on the big screen, some want to do both. In order to account for these potential differences, I try to obtain segmentation of behaviors which could tell different engagement patterns and going one step further I want to understand where to focus in making it a stickier experience for each user.

The foundation of my approaches is building out an engagement feature bank. I used Python Spark shell and SQL to pull data from HDFS database and complete feature engineering. For instance, I define 42 metrics that are relevant to user engagement and randomly pick a sample of

over 500,000 records to understand major patterns of how users engage with the app, which means how users behave similarly or differently with each other.

To obtain segmentations, I used unsupervised learning method clustering. Before clustering, I did some pre steps to help reduce the noise in data. For features that have highly skewed distributions, I used log transformation to preprocess the data. Dimension reduction helps reduce the curse of dimensionality while doing segmentations. Specifically, I used PCA to reduce the dimension and determined the number of principle components by setting a cut-off of 80% explained variance.

For the segmentation part, I used K-means clustering algorithm and choose 5 clusters to segment by the elbow method. After clustering on reduced dimension data, we obtain 5 groups different in size. To have an overview of each cluster, I also defined a target variable to help measure their performances, indicating whether users come back and play in the next 7 days.

To figure out what makes clusters unique, I selected out top 5 features that contribute to the most cross-cluster variance. Then, to get a comprehensive view of the clusters, I extended the 5 features to 15 based on the feature categories. After examining from both performance and feature perspective, I could name all 5 groups of users and identify engagement patterns of my sample data.

As the next step, I tried to understand the relationship between user behavior patterns and whether they come back and play at least once in the next 7 days. I chose several binary classification algorithms to predict viewership retention including SVM, Logistic Regression,

Random Forest Classification, XGBoost Classification, KNN Classification and decision tree, and pick 4 of them with relatively higher benchmark performance for further fine-tuning and evaluation.

Specifically, in Python, I used random search to grab a range of parameters and further used grid search to obtain the hyper parameters of best model for each algorithm. With tuned models, I used three metrics to evaluate each model including Accuracy, F1 score and area under ROC curve. XGBoost Classification model has the best overall performance. With its prediction, I select top features by importance to see which features has the greatest impact on the outcomes.

From segmentation and modeling prediction, I understood what a sticky user looks like but how can we move everyone towards that direction? What actions should we take first? Can we even personalize the recommendation? One way to answer previous question is to interpret the model I built. I used a method called Shapley value based on cooperative game theory. In short it explains why the model thinks a user is less vs more likely to return and play compared to the average. Shapely value plot tells features' positive and negative effects and their magnitudes from which we could give inferenced personalized recommendation to improve the engagement.

Above all, I figured out representative engagement patterns from clustering and segmentation, key features encouraging a stickier user experience from modeling and target personalized recommendations from individual level. If I could have more time, with potential inference on features, I need to conduct A/B Test to further prove the causation in order to make product feature improvements. What worth noticing is that engagement and behavior patterns could

change over time. Also, with potential improvements in product features, we may check on a regular basis to adjust the model and therefore to complete a development circle. I can also improve from data perspective, for example extended demography and longer time period may give a more comprehensive view.

Beyond my project, the overall experience acquired through this internship can be summarized into three aspects: Teamwork, Innovation and Results. As an intern, I worked mainly on my own project but everyone in the team is so supportive that helped me solving problems in technical and logistics. I enjoyed conversations with colleagues within or cross teams and hearing their Apple experience which could bring me inspirations and let me know more about the company culture which made my internship a more rich and comprehensive experience. Solving Data Science problem is like climbing. There's never only one way to get to the destination. During my internship, I not only implemented traditional theories and tools that I learned from school courses, but also learned and practiced some new methodologies. The process of research is a kind of innovation and that's what I want to achieve in my future study and career life. Finally, if speaking of result, I did a quite successful presentation and refined my presentation skills, more specifically, I learned about how to draw high-level conclusions and deliver insights to all kinds of audiences. All in all, what I harvested through this internship are far beyond my imagination and will benefit my future career life significantly.

Internship completion and approved by: 9/1/2020

Employer/ Manager Signature: