Blue Pandas - Asurion Phone Claims Forecasting Meeting Notes

Wednesday, February 15th at 11:30 am (In-Class Sprint Meeting)

- Discussed whether we want to use external datasets
    - Each team member is assigned a specific external dataset research issue
    - Deadline is during Monday's meeting as Tiffany needs to submit external dataset information on BrightSpace by Monday night
- Since the team is waiting on Asurion dataset to get posted, we will each look into different time series models as a knowledge share to discuss during the meeting on Monday

Monday, February 20th at 2:30 pm (Backlog/Scrum Meeting)

- Gave a quick overview of the ACCRE recitation tutorial
    - Clarified how we are supposed to use ACCRE and GitHub
    - George is to send out his notebook to ensure everyone on the team has the same data file access
- George already verified the data
- Tiffany showed the iPhone external dataset research from statista
    - Datasets include market share, sales share of the Apple iPhone by model, and iPhone unit sales
        - Tiffany is to reach out to Yingxiao to confirm if the Asurion dataset is only on US data or if it's all of their global data
    - Team confirmed to use external datasets if applicable
    - Team discussed potential problems with using quarterly data since Asurion's data is weekly
- Team discussed modeling research - autoregression, weighted moving average, exponential smoothing, ARIMA and SARIMA, box-jenkins, and Facebook's Prophet
- Brainstormed EDA ideas to have in our backlog
    - Each team member will create 2 specific EDA issues from the large EDA Brainstorm issue's checklist
        - We will assign the specific EDA issues during Wednesday's sprint meeting

Wednesday, February 22nd at 11:30 am (In-Class Sprint Meeting)

- This is the start of Tiffany's week 2 as team leader
- Checked in on the health of the team:
    - Communication is overall still going well and no one is feeling overwhelmed from last week's research tasks
    - Everyone completed their tasks as indicated in the last meeting
- Confirmed with Dr. Zhang that we can use the external datasets at anytime

- Tiffany has not yet heard back from Yingxiao. As advice from Dr. Zhang, we will wait a few more days before reaching out again or we can go with Dr. Zhang's assumption that the Asurion data consists of only US claims data.
- Team reviewed and finalized the backlog EDA tasks discussed from the last meeting and assigned 2 tasks for each person in the sprint
    - Besides George's tasks due to needing his items before everyone else can start, all of the EDA tasks are due by Monday for a review during the team meeting
    - Discussed expectations and clarified meanings of quarterly data, fiscal year, seasons, and how we want to look at data across time
    - Clarified and placed the clustering issue into the backlog
- Advice from Dr. Zhang is to use the kanban board to assign issues instead of going to issue tab on GitHub
- Advice from Dr. Zhang is to keep notes on knowledge share. The team agreed that we will use the meeting notes as a way to keep track of knowledge share.
- We discussed how we will name our EDA notebooks for consistency. For example, Kit will label his notebook as 21 and Tiffany as 22.
- Team discussed seasonality aspects
- We will do a one-time shift of having the Monday meeting to be at 4:30 pm

Monday, February 27th at 4:30 pm (Backlog/Scrum Meeting)

- Team discussed how to complete assignment #3
- Kit, Sruthi, and Tiffany went over their EDA notebooks
    - We noticed that there are many outliers in the data. For the next sprint, we will subset the data to remove older iPhone models.
    - Sruthi showed that there were low correlations among the different phone features.
    - Kit showed how Apple goes through seasons of peaks and brought up interesting points for the team to look into next sprint. He showed how Apple's iPhone claims have gone up tremendously last year, and how Samsung phone claims have been generally consistent over time, which shows how Samsung doesn't go through much seasonality.
    - Tiffany showed the feature engineering for seasons, quarters, and holidays. We saw that there is an imbalance in the number of rows for each category based on our data ranges. For the next sprint, look to see how phone models have been categorized by the specific labels.
- George parsed out the data file to fix a formatting error and Tiffany will output a new data file that has the seasons, quarters, and holidays included on ACCRE.

Wednesday, March 1st at 11:30 am (In-Class Sprint Meeting)

- Kit is the team leader
    - Team health check - everything is good with the team and no communication breakdowns
    - We expect that we're going to have limited capacity this coming week

- We formalized the issues we had on backlog with making sure we do EDA with a subset version of iPhone model data
- We discussed what potential EDA we can do with the external datasets
- We plan to have a quick meeting on Sunday to discuss what we want to cover in our client visit

Sunday, March 5th at 2:00 pm (Meeting)

- Plan to discuss about data cleaning, the time-series analysis' rolling mean, and the feature engineering on season, quarter, and holiday
- If there are follow-up questions from the client, we plan on going onto GitHub to show our notebooks

Monday, March 6th at 11:30 am (Client Meeting #1)

- We asked the client the question about whether we should predict iPhone 14 in general or should we break it down with having iPhone 14 Pro Max.
    - The client said that we can just categorize all of the data as iPhone 14. We also noticed that there aren't a lot of iPhone 14 Pro Max, so we will just categorize all of it under the iPhone 14.
- We will work on our EDA part II over spring break due to our recent workload. We will also look into a bit on researching modeling during the break or right after the break.
    - Everyone is continuing on their EDA tasks from the last sprint but with the subsetted iPhone data.
    - George will work on random forest and EDA on the external datasets. He will look at the external datasets to see if we can use any of this information.
    - We will need to look into ARIMA and SARIMA models. Sruthi and Tiffany will look into the different models to figure out which one they want to pick to use. We will use these autoregression models.
    - Surthi wants to look into XGBoost to see if we can use it for our modeling phase.

Wednesday, March 1st at 11:30 am (In-Class Sprint Meeting)

- Discussed the issues we will cover over Spring Break
    - Due to their availability, Tiffany and George will connect over the last few days of Spring Break to touch-base on their work and do any peer reviews
    - George and Tiffany will work on EDA on external dataset, new dataset, and work on some modeling
- Clarified with how we are going to include the rolling mean average after subsetting the iPhone data as a feature in the dataset
- We clarified that we don't think clustering through EDA would apply to our dataset since we are doing supervised learning
- We double checked our team charter to make sure our team leader schedule is still consistent. George will start being the team leader in our next sprint team meeting after spring break.

- We plan to have our feature set finalized by Monday, 3/20 during our backlog/scrum meeting. Though, we still plan to iterate through the data science workflow based on the info we learn from the model.

Wednesday, March 8th at 11:30 am (In-Class Sprint Meeting)

- Everyone continues to work on their EDA or feature engineering issues over the break
    - George is to work on EDA for the external dataset
    - Tiffany is to make a new EDA notebook that breaks down by iPhone models based on seasonality, quarter, and holiday
    - Sruthi is to make a new EDA notebook that breaks down iPhone models based on their initial features
    - Kit is to do feature engineering to create iPhone model group labels - we will focus on the relatively newer iPhone model releases
- We will start modeling after the break, but we can take some time during Spring Break to look over some models
- We also reconfirmed what is our deadline to finish modeling once we come back from Spring Break

Wednesday, March 22nd at 11:30 am (In-Class Sprint Meeting)

- Officially based on our charter, we switched the team leader role to George, but the team discussed the initial confusion as Kit and George's weeks overlapped with Spring break.
- Kit, George, and Tiffany discussed the issues they completed over Spring Break
    - Kit showed the revised dataset with the phone model labels
        - The team discussed including labels for iPhone SE since iPhone generation II and III are considered relatively new releases.
    - George showed his EDA on the external dataset
        - The team agreed that the external dataset can be used as a final consideration based on our model's output. For example, we can use the information from the external dataset to make assumptions, such as include a 0.02% growth rate for all iPhone claims data based on the growth in Apple market share.
    - Tiffany showed her new EDA notebook that differentiated different iPhone models based on seasonality, quarter, and holiday
- Team assigned and created issues about the models to look into as we start the modeling phase:
    - George will do Random Forest
    - Kit will do XGBoost
    - Tiffany will do ARIMA
    - Sruthi will do LSTM

Monday, March 27th at 2:30 pm (Backlog/Scrum Meeting)

- Kit, George, and Tiffany showed the results of their initial models and some problems they have been encountering.
    - Team agreed the importance of aggregating the claims by ‘weeks_monday’
    -
- George and Tiffany plan to visit the Professor to ask about how to conduct predictions with iPhone 14 data, questions about ARIMA, and questions about Random Forest

Wednesday, March 29th at 11:30 am (In-Class Sprint Meeting)

- Due to confusion, we allowed George to remain team leader for this meeting, but we all agreed that this is George’s last day as team leader.
- Clarified with the Professor that we will be listing Kit as the team leader for the ICA submission as we will have two more ICAs afterwards
- Kit, George, and Tiffany showed updated results of their models from the last meeting and work on model tuning
    - Kit will conduct XGBoost model tuning by utilizing hyperparameters
    - George will conduct Random Forest model tuning by utilizing hyperparameters
    - Tiffany received feedback to try normalizing the data and finding other hyperparameters for ARIMA given how one of the currently created models did not make any real-world sense
    - Tiffany brought up about potentially moving to linear regression model once ARIMA model is more finalized, but she still plans on finding alternative ways to work on ARIMA
        - The linear regression model will be worked on if there is time allowed
- Team discussed RMSE, MSE, and WMAPE to confirm about the error we need to report to the client, which is WMAPE

Monday, April 3rd at 11:30 am (Client Meeting #2)

- Confirmed with client to remove April 13, 2023 datapoint as the data compiled was potentially incomplete based on when the data was pulled, this leaves the team with 85 weeks of data
- Confirmed with client about how to do WMAPE matrix if we are not splitting up by phone color and phone size
- Kit, George, and Tiffany showed Yingxiao XGBoost, Random Forest, and ARIMA models
- Yingxiao provided feedback about XGBoost and Random Forest, and she brought up as reminders that some iPhone models, such as iPhone 8 are in the declining product stage
- Yingxiao brought up about creating an ensemble with XGBoost and ARIMA
    - Yingxiao also brought up doing a few weeks' lag to predict the future. This was the discussion drawn out on paper.
- Team clarified the discussion on how to calculate WMAPE and how there will be a function passed around to make sure everyone remains standardized

- Team clarified with the Professor that we should have WMAPE depend on our current data. If we do get the new iPhone 14 data, then we can do WMAPE based on the new provided iPhone 14 data.

Monday, April 3rd at 2:30 pm (Backlog/Scrum Meeting)

- Everyone provided a recap and updates about all of their models and additional logistical information so that everyone is up to date
    - Everyone is to continue working on their current models and we'll discuss any new developments on Wednesday before determining next steps
    - Team decided to currently not conduct the ensemble of XGBoost and ARIMA and to currently focus on fine-tuning the original models
    - Sruthi provided the update that she is working on LSTM and she is encountering difficulties with the model. She is getting modeling help from a friend outside the program.
    - Everyone agreed that we need to figure out what's the best model and use it to report in our final presentation as the main focus while the other models are used to show that we attempted to look at different models
- Discussed whether features for tree-based models (Random Forest & XGBoost) should utilize the same features or if they can be different.
    - Team concluded that we need to make sure that we utilize the same iPhone 14 test dataset to compute WMAPE so that we can compare different models
- Team discussed the timeline:
    - April 15th - Team needs to completely stop modeling and make everything finalized
    - April 16th - Team works on presentation and the team needs to get together to do one practice run that is separate from the optional in-class practice run
    - April 17th - Team is assuming our optional in-class practice run happens on that day
    - April 19th - Presentation

Wednesday, April 5th at 11:30 am (In-Class Sprint Meeting)

- Everyone gave updates on their models and the work they need to do next week.
    - Sruthi is working on getting LSTM running
    - Tiffany will be moving on to linear regression as a few ARIMA models have been built
    - George discussed about the random forest and some things he encountered in his code
    - Kit discussed about XGBoost, created a WMAPE, and the general scores we're encountering on WMAPE
- Team confirmed to have a practice in-class rehearsal on April 17th

Monday, April 10th at 2:30 pm (Backlog/Scrum Meeting)

- Kit, George, and Tiffany provided updates about their models. Each person is still working on their models.

- Team discussed the general timeline for the rest of the project
- Kit discussed how we need to scale our predictions by a factor to provide an accurate prediction for iPhone 14. Kit is working on providing the factor scaling based on each team member's dataset that was used for that particular model.
- Kit is also working on a script to help everyone be able to indicate what are their predictions for iPhone 14
- Team discussed the presentation flow - we will list out our WMAPE and we will also include each model's prediction values
- Team discussed the issue with the "Asurion_data_additional" file that was provided last week.
    - Tiffany discussed how iPhone 14 data is inconsistent compare to our original data file
        - Inconsistent as the file doesn't include iPhone 14 pro max and there are missing information with the storage size options (for example, the iPhone 14 data doesn't include 512 gb)
    - Team decided to not use the new "Asurion_data_additional" file as we determined that the inconsistency may create further problems with our modeling, due to time constraints, and we don't think the additional data points will help our model significantly
    - Team has also determined it would be too difficult to impute the data as we don't have a good base reference line for iPhone 14 pro max claims.

Wednesday, April 12th at 11:30 am (In-Class Sprint Meeting)

- Team agreed that we'll all generally stop modeling unless anyone has any currently unfinished models
- Team discussed the difference in WMAPE values for each of the models
- Team has agreed that we'll use Random Forest as the main focus of our presentation
- Team checked the new "Asurion_data_additional" file that was provided again and noticed that it is correct and most inconsistencies are gone
    - We decided to not use this dataset for all of the models and just only use this dataset for the Random Forest model
- Kit will work on creating a data file that works for the Random Forest model

Friday, April 14th at 11:00 am (Presentation Meeting)

- Team discussed how we wanted the presentation to flow and who is speaking which parts of the presentation
- Everyone must have their slides completed by Sunday's meeting

Sunday, April 16th at 4 pm (Presentation Meeting)

- Team discussed the slides that were created and provided feedback on content
- Discussed what material we should cover in the presentation

Monday, April 17th at 2:30 pm (Backlog/Scrum Meeting)

- Team discussed the feedback covered in the presentation dry-run with Professor Zhang

    - Made slide edits to incorporate additional text on graphics
    - Team decided to include an Appendix section to provide further details on modeling due to Professor’s feedback on needing to add additional information on the models in our slides
        - We plan to have the Appendix slides be more in depth so that we can directly easily refer to them if the case arises during the presentation
- Team will try to incorporate standardize features depending on everyone’s capacity
    - We don’t plan on having it perfectly standardized
    - George and Tiffany will meet separately to discuss what features that can be used in Random Forest and Linear Regression if possible
- Team will meet on Tuesday to go over presentation again
- Every team member will work on the technical report to be due on Thursday so that it can go into editing
    - Two team members will handle report editing while the other two team members do GitHub repo clean-up

Tuesday, April 18th (Sprint Meeting)

- Team discussed the newly added appendix slides and conducted a dry-run
- Discussed the new features added in the Random Forest and Linear Regression as an attempt to have feature standardization
- Discussed how we wanted the technical report to flow
- Worked on cleaning the GitHub repo