CSci 5123 Replication Assignment Cover Sheet

Project Team (list all members):

1. **Luke Strassburg <stras134>**
2.
3.

DOI/URL for Paper:

- https://variancejournal.org/article/31370-recommender-systems-for-insurance-marketing

Checklist for Replication Assignment Grading

___ Basic Replication

- Was code available? (yes/no)
    - If yes, what was needed to get it to execute?
    - If no, what libraries/tools did you use to replicate?

Yes and no. There was code for the 4 Machine Learning models (CTB, DL, LGB, and XGB) which gave me an idea of how they performed their analysis. All that was required for these was downloading the required packages and updating their documentation to fit the modern versions of them. I also had to create the features.csv which is the input to the models because it was missing from the provided GitHub repository. I think I duplicated it as close to as possible to what they had, but I definitely have a few things differently.

However, there was a lot of code missing for this project that I had to make myself. I wrote all the code for Association Rules, Item Based Collaborative Filtering, LIBFM, GLRM, ALS, SLIM, and GLM by myself, using the other 4 as a rough template. This required using various packages mentioned in the original paper across both Python and R. These files are found within my zip drive. Check the README for which files hold the code for which models.

- Was data available? (**yes**/no)
    - What steps needed to find/adapt data?
- **Yes data was available. It was a 10000 policy subset of the 10 million used in actual analysis. I had to make the features.csv used to filter the predictors and information within the dataset.**
- Did basic replication produce expected results? (yes/no)
    - Was parameter exploration/tuning needed?
    - What steps taken to get as close as possible to expected results?
- **Results are decently similar. Some good variation but that could be due to the different underlying datasets since I only had a 10000 row subset of the original 10 million rows.**
- How many algorithms, datasets, metrics were included in basic replication?
    - **11 algorithms (All the ones mentioned above)**
    - **1 Dataset (10000 row subset)**

    - **Metrics (AUC Only)**

Yes Enhancements

- Additional datasets (list ones evaluated)
- Additional algorithms or variants (list ones evaluated)
- Parameter exploration / tuning (list)
- Additional metrics (list ones included)
    - **I added Accuracy, Precision, Recall, Pr-AUC, F2-Score for all the models. Looking for identifying more positive predictions to identify potential sales opportunities.**
- Additional forms of complexity added or encountered

Yes Written in form of a Replication Paper (yes/no)

Yes Other factors that should be taken into account

I was originally at Replication Phase 2, planning on creating another set of recommendation models trying to model for the bought parameters within the dataset (Modeling based on future purchases rather than current ownership). However, I ended up spending much more time reimplementing all the models without given code than I anticipated, so I decided to cut it out.