Mid-project Review

Startup funding predictor

MSIA 432
Finn Qiao

Highlights

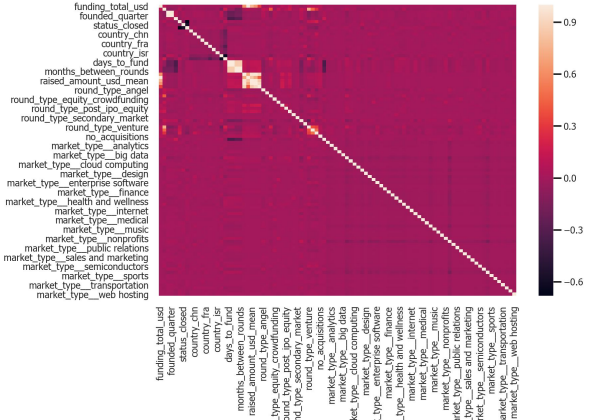
1. Understanding how the data linked across multiple datasets was difficult and it was challenging having to cherry pick and generate new features that were actually relevant.
2. Entity recognition is very difficult and was incredibly challenging with only NLP analysis but managed to whittle features down by combining some metrics.
3. Feature selection with Recursive Feature Elimination from 105 to 30 features.

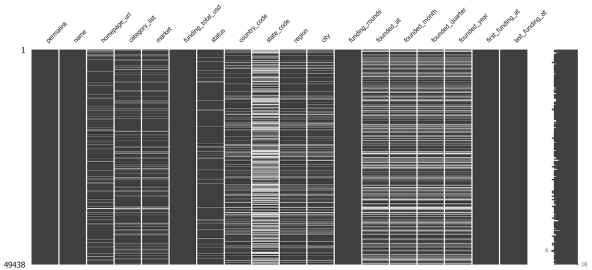
Progress

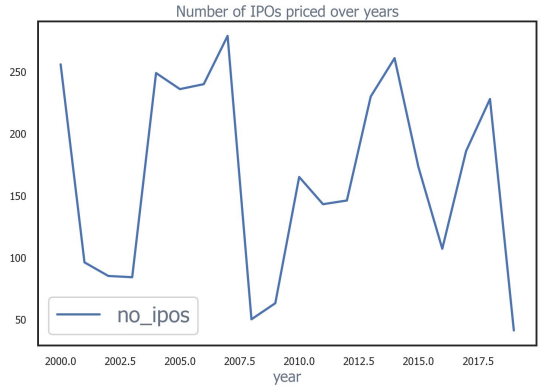
1. All data now cleaned and new features all generated.
2. Aggregated dataframe created and persisted to S3.
3. Recursive Feature Elimination run on extra trees regressor to select top 30 features.
4. RDS schemas set-up to store top 30 features and user interactions with app.

Analysis

1. Incorporated past IPO trends into data.
2. Visualized correlation across variables.
3. EDA in conjunction with entity recognition to collapse all market and industry types to top 50.







Lessons Learned

1. Difficult to parse out trends within each market category as they overlap too much.
2. Entity recognition requires way more examples and better algorithms rather than simply NLP.
3. Recursive feature elimination with 10 fold CV is way too computationally expensive and does not guarantee better or interpretable results.

Recommendations

1. Next steps include further tuning of feature selection.
2. Ingesting of final filtered data to RDS schema for faster querying.
3. Finalizing of the user input needed to generate inputs into model.
4. Scripts that run extra trees regressor and obtains predictions.
5. Base flask app that takes in user input.