glassdoor

DS Intern Assignment – Mitali Bharali
Graduate Student – UT Dallas
Major: Data Science
mitali11bharali@gmail.com, 4693803996

Analysis Workflow

Data Cleaning

- Remove NaN, duplicates
- Format columns, data types

Data Visualization (tableau)

- Variations across variables
- Outliers
- Correlation between columns

Data Modelling

- Supervised Learning – Classification models

Ensemble Learning

Model Evaluation, Boosting (python)

- Bagging, Pasting
- Adaboost, Gradient
- Cross Validation and GridSearchCV

Performance and Feature selection

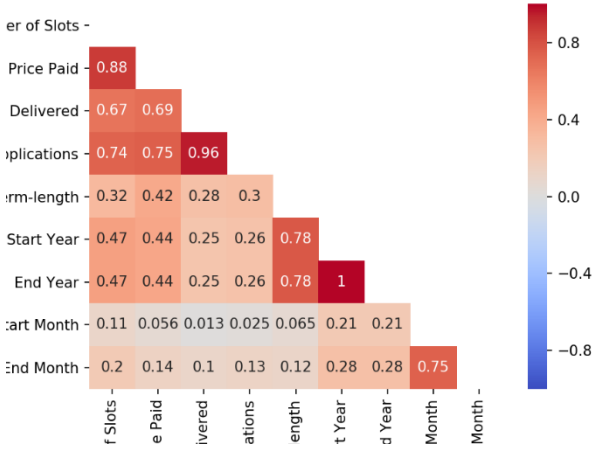
- Accuracy, Precision, Recall and f-1 Score
- Random Forest Feature selection

Data Cleaning

	Employer ID	Number of Slots	Price Paid	Marketplace Value Delivered	Applications	Renewed?	term-length	Start Year	End Year	Start Month	End Month
0	40	125	67125.0	147711.043445	9647.7	0	393	2015	2016	2	2
1	43	100	85025.0	62214.929769	7182.2	1	365	2015	2016	12	12
2	72	125	91290.0	59242.473666	6571.3	0	730	4031	4033	4	4
3	94	35	21480.0	27519.574564	2558.8	1	365	2015	2016	12	12
4	102	50	26850.0	24337.924740	1501.1	1	365	2015	2016	5	5

- Term length: calculated with start date and end date
- Date and month are extracted
- One hot encoding for locations if needed

- Heatmap generated for correlation judgment among various columns, usually drop one of the columns that's highly correlated



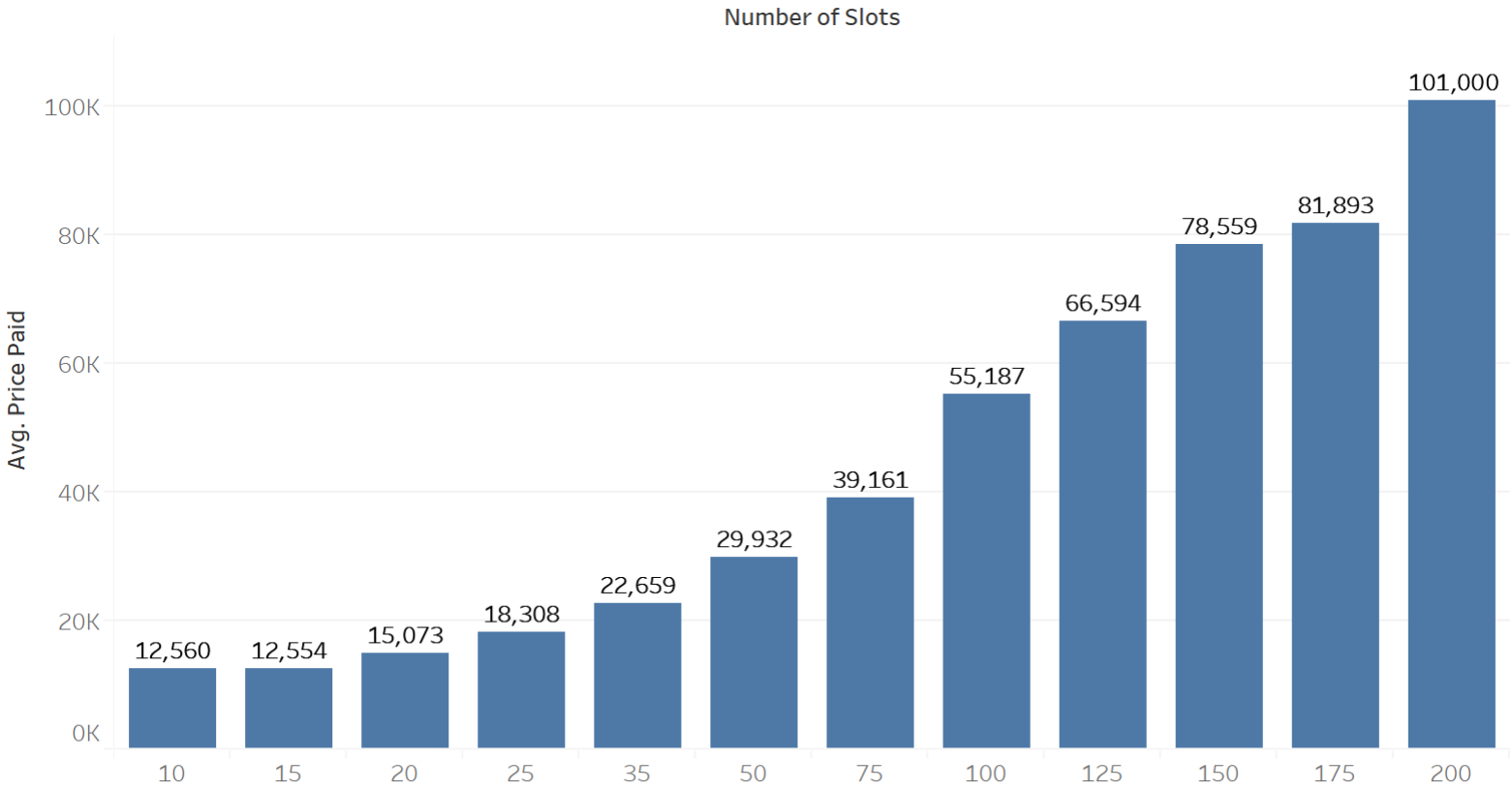
Employer ID	Employer City	Employer State	Price Paid	term-length
40	Belmont	California	67125.0	393 days
43	Northbrook	Illinois	85025.0	365 days
72	New York	New York State	71600.0	365 days
	na	na	19690.0	365 days
94	Houston	Texas	21480.0	365 days

There are duplicate Employer IDs, hence to understand customer behavior, treat each customer uniquely

1)Variations across Job slot packages

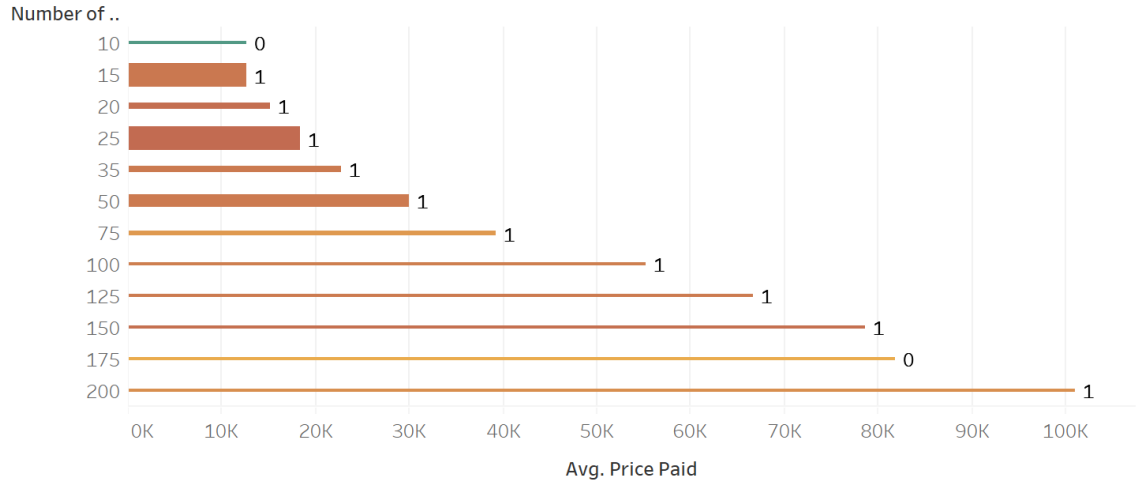
Using Tableau

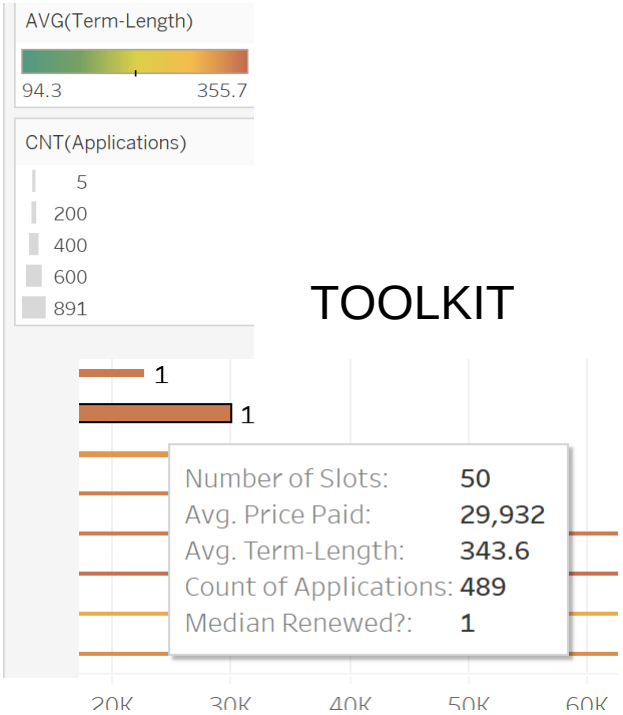
Prices over different no of slots



Price Distribution

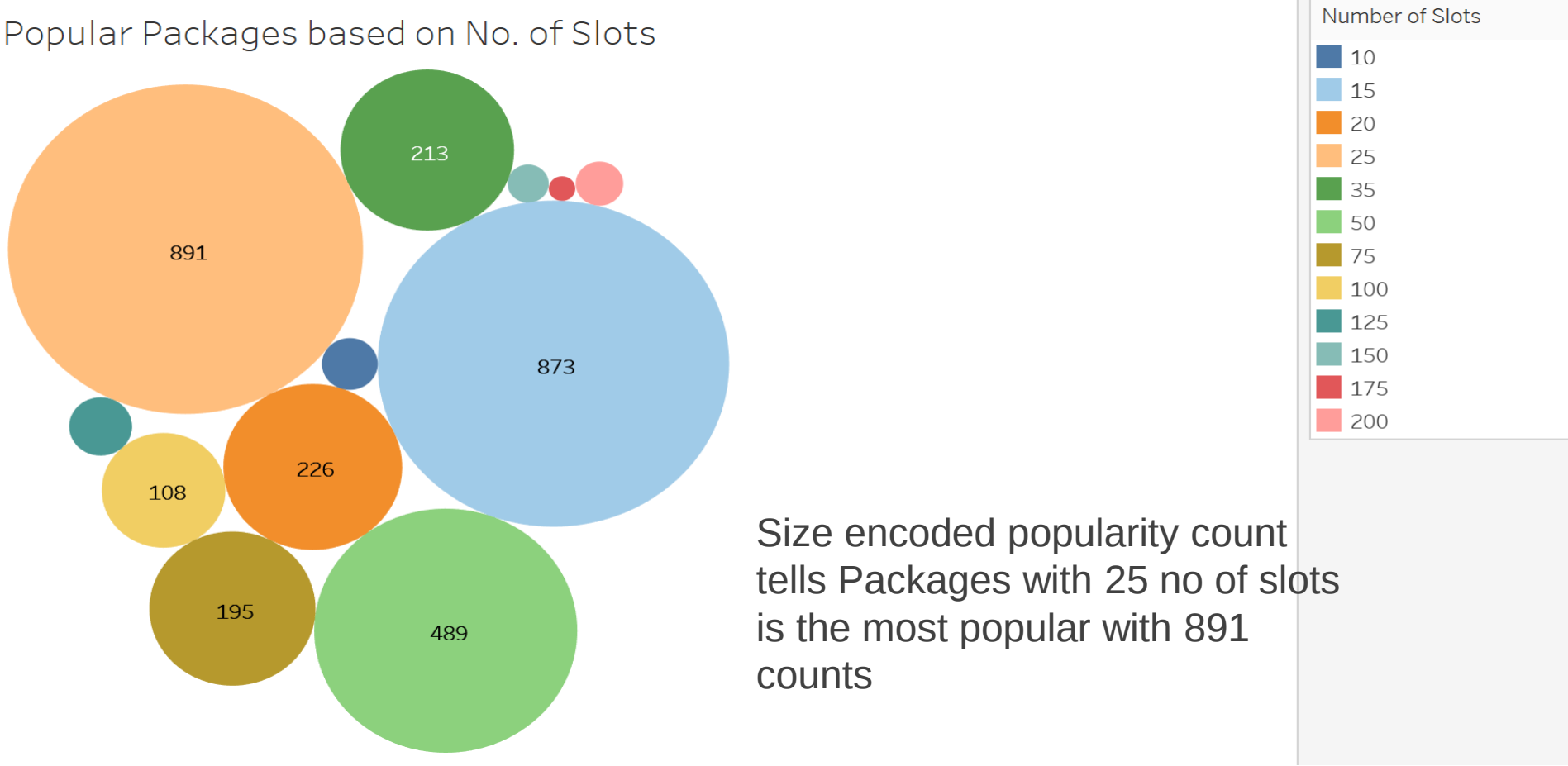
Term Length, Slot Package, Avg Price vs Renual Performance, count no of applicants recieved





Price & Term vs others

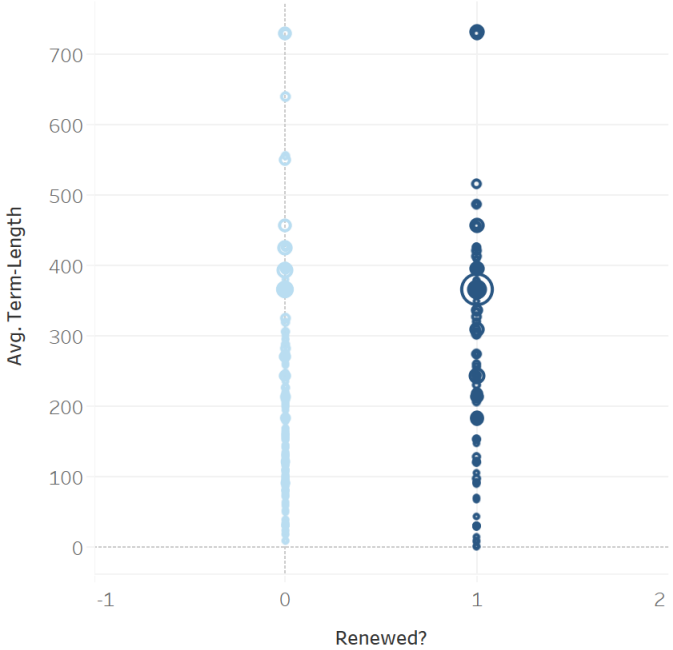
Even when more price is paid for higher slots, count of applicants received (encoded with size) is lesser. The variation of term length is also encoded with color, telling slot of 10 was for less duration and never renewed and also didn't obviously receive much applicants



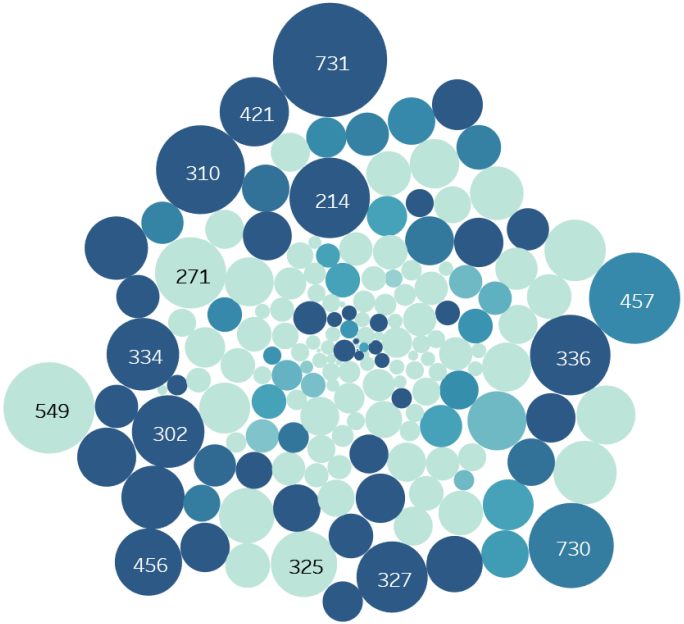
Size encoded popularity count tells Packages with 25 no of slots is the most popular with 891 counts

Popular Packages

Term Length vs market value and renual



Term length vs no of applicants with renual



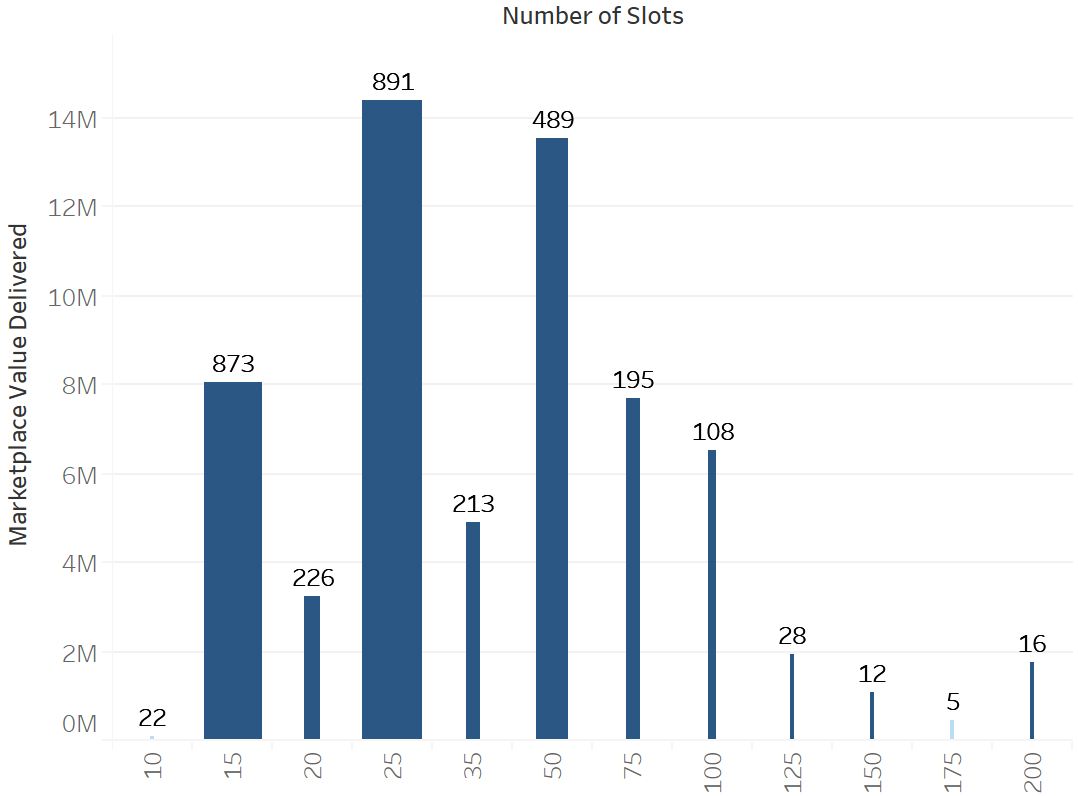
Term Length

Variations

Left graph : market value delivery is almost as evenly distributed with term length
Right Graph: Size tells us the median no of applicants against the label of term length; dark color is renewed and light is not

2) Variations in Delivery performance

Market Value Delivered vs slots, renewal across customers



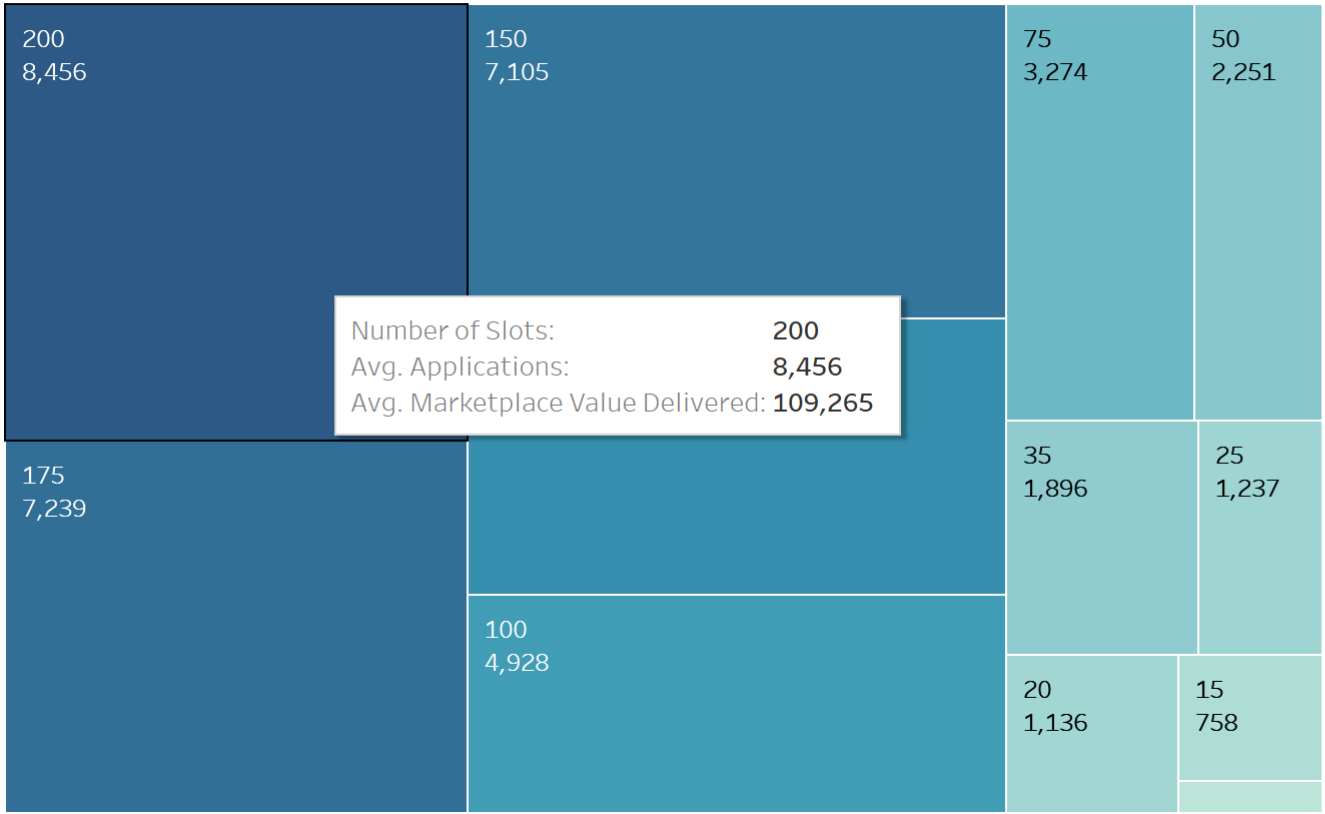
MEDIAN(Renewed?)

0.000 1.000

CNT(Employer ID)

5
200
400
600
891

Market place value vs applications



AVG(Marketplace Valu...

3,132 109,265

2) Retention Analysis

Used Python

Data Modelling

Machine Learning

Data Models

Logistic, SVC(kernel), Decision Tree

- Best Accuracy for decision tree – 74% test
- Easy to change colors, photos and Text.
- Model Evaluation using GridSearchCV; tuning hyper-parameters important

Ensemble Learning

Programmer

Divides dataset into several set, more accurate

Model Performance

Programmer

Accuracy, but Precision recall more for classification models

- **Precision** is most important in this case
- Need to predict renewed (1) correctly, reduce false positive values more, hence precision

3) Feature Selection

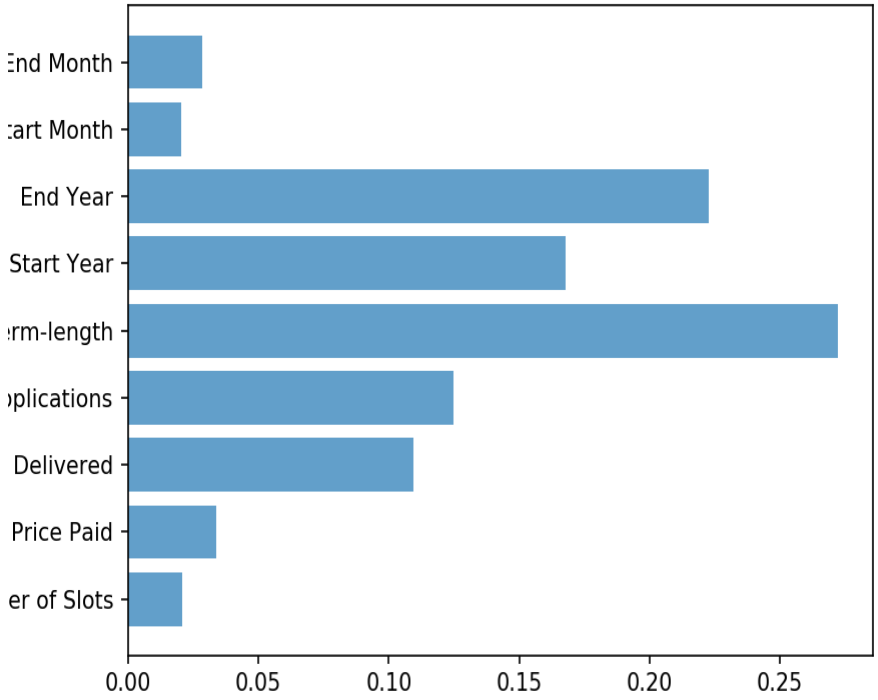
Factors that have the greatest impact on retention | Predicted through Random Forest feature selection

01 TERM LENGTH: is the most important predicting value

02 Start, End – Years are prominent

03 Applications are next

04 Marketplace Value Delivered



3) Feature Selection

Removing correlated columns from the models, we receive different results

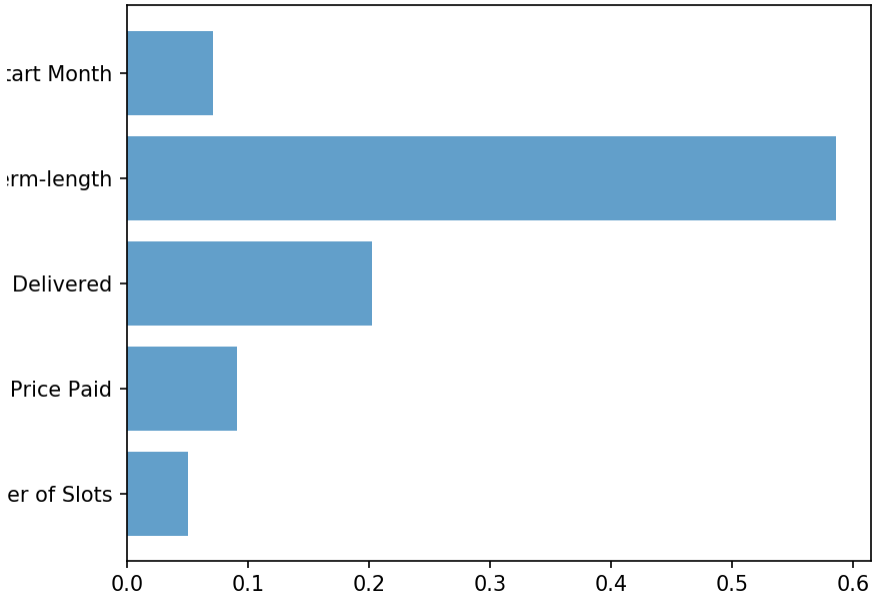
01 TERM LENGTH: is the most important predicting value

02 Market Value delivered

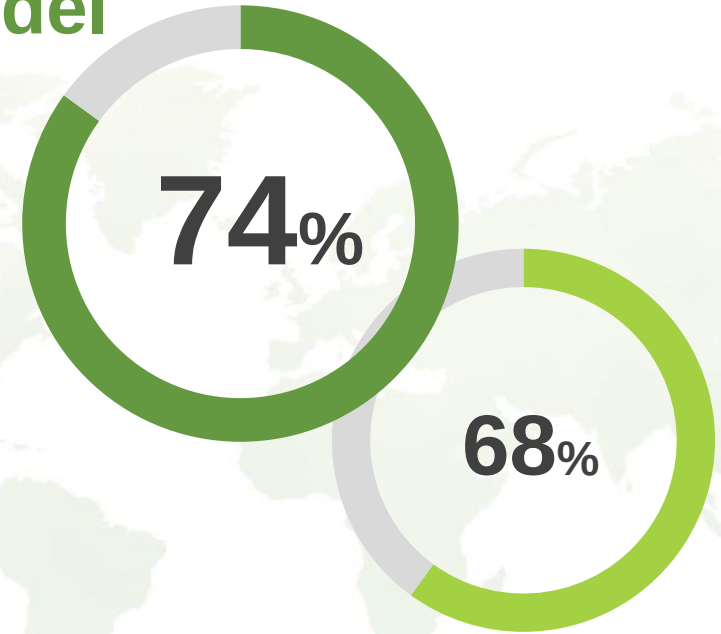
03 Price Paid

Feature Selection

```
In [47]: forest.feature_importances_
Out[47]: array([0.05019126, 0.09074215, 0.20268423, 0.58549846, 0.0708839 ])
```



4) How well does the Prediction Model work?



74% Precision Score
We need to reduce False negative (type 1) error; recognize more of renewed = 1

Tuned Parameters to –
n_estimators=500,
max_leaf_nodes=16

For classification models, more than accuracy, model performance through precision recall matters

68% Accuracy - Adaboosting on Decision Tree

Max Depth – 2
n_estimators = 500
Learning rate = 2

SVC

	precision	recall	f1-score	support
0	0.87	0.17	0.28	237
1	0.65	0.93	0.76	402
avg / total	0.72	0.66	0.60	690

Decision Tree

	precision	recall	f1-score	support
0	0.88	0.19	0.31	237
1	0.66	0.92	0.77	402
avg / total	0.74	0.68	0.61	690

Model Performance

Accuracy, Precision, Recall and F1-scores

Random Forest

	precision	recall	f1-score	support
0	0.78	0.27	0.40	237
1	0.67	0.90	0.77	402
avg / total	0.71	0.68	0.64	690

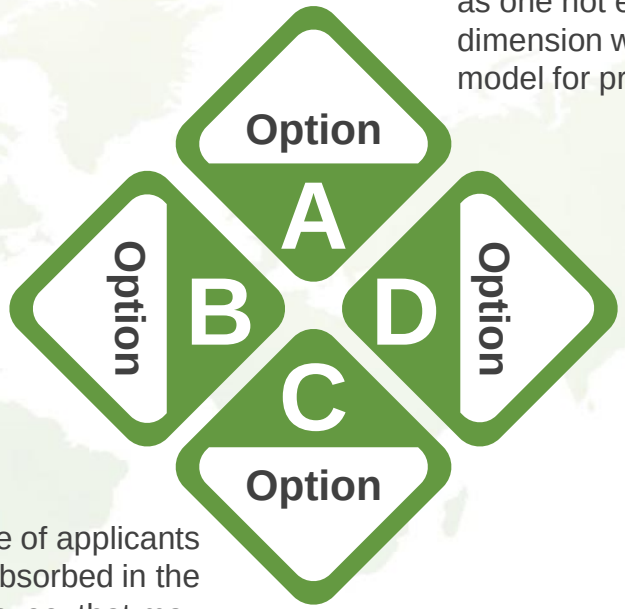
Adaboosting

	precision	recall	f1-score	support
0	0.79	0.25	0.38	237
1	0.67	0.90	0.77	402
avg / total	0.70	0.67	0.63	690

Improve analysis

4) What other factors might you want to investigate to see if they could improve your analysis?

Use location, (either City or State) as one hot encoded → then reduce dimension with PCA to use in the model for prediction



Heat map can tell us correlations and removing correlated columns can give better predictions

Data is biased, since most of it is of renewed status 1, very less knowledge of renewed status 0, hence collect more data to remove imbalance

If there is knowledge of applicants converted/absorbed in the company as employee, that may give more credential as a value delivered

5)Recommendations

Based on your analysis, what modifications would you recommend we make to our ad platform algorithm to improve retention

Term length variations over slots tell which slots to concentrate,
Term length over price also tells, the optimum price for every term and hence a more renewal chance

Company should concentrate more on slots of 50, 25, as they show more no of applicants, term length remains longer.
They should drop less no of slots package as they have no renewal and no term length or applicants.

More data should be collected to predict price better, more no of records, and all also more features

Thank you