

Travelers Case Competition 2017

KANGAROO AUTO INSURANCE CHURN PROBLEM

TEAM – CHN@UCONN

Identifying the indicators that will help predict if a customer will churn from the company or not

Background

Kangaroo is an Auto Insurance Company based out of Australia offering Property insurance policies, Auto insurance policies etc. to customers across the globe. The given dataset provides data on customers from the US market.

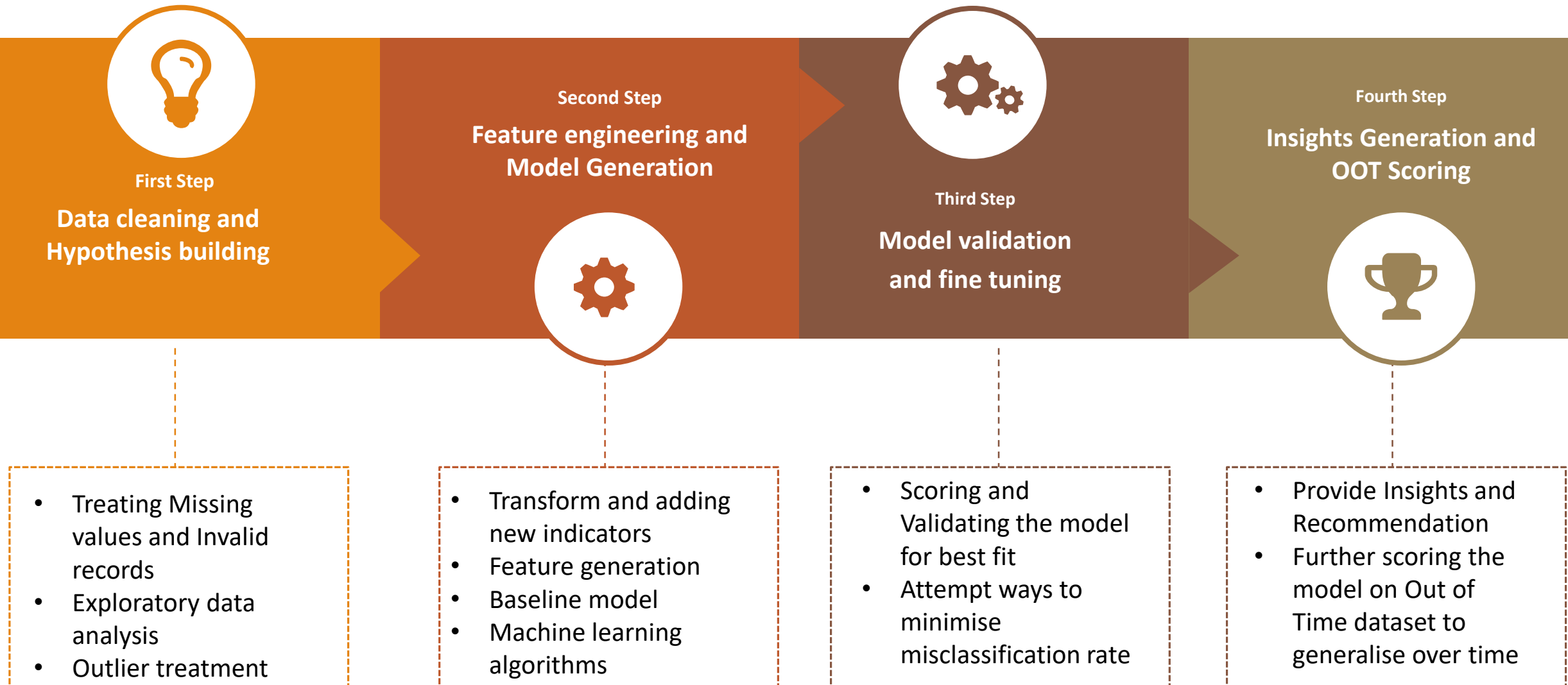
Problem Statement

Around 7K Property Policies has been provided for the years between 2013-2016 out of which ~26% were cancelled. Using statistical evidence, our aim is to identify the pattern of customers who are most likely to cancel before the end of the next term (2017)

Data availability

Two datasets- Train and Test were provided. The training dataset contains 7K records and the test contains 2K records with 16 indicators. Some of the indicators are the gender of policyholder, tenure with kangaroo, # of children and adults living in the property, etc. are given.

Steps carried out for factor analysis of churned customers vs non churned customers



Exploratory data analysis was conducted to come up with high level insights about the business

Data Cleaning - Train



Missing Rows	Age >76	Churn -1
2%	1%	.03%

~ 3% data improper – Decision to remove

Data Cleaning - Test



Missing Rows	KNN Imputation
Age >76	Value Flooring

Business Deductions



Customers / year



3/4

Retention rate



72%
Married



53% 47%
Gender



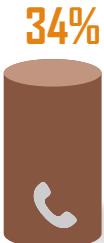
20%
Claims



Broker



Online

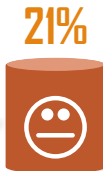


Phone

Sales Channel



High



Medium



Low

Credit score

Feature engineering proved to be important in a dataset with minimum variables.
Additional features will be useful for different models

Variable enhancement

- Categorical variables were converted to indicator variables in order to study if there are interaction effects. For ex. The effect of a particular variable occurring more in a year
- Math transformation was used to get best representation of features
- Additional features were extracted as well

Year 2014 * Claims

+


Tanh Square Cube Log Exponential

+


Tenure/Len at residence = CLTV indicator


Geospatial analysis


- Connecting the customer zip code to the database helped linking the state and city which is easier to model when compared to Zip code
- Building on a hypothesis that a zip or city that has lot of cancelations happens due to negative perception of customer; we create indicators for each zip and city



+



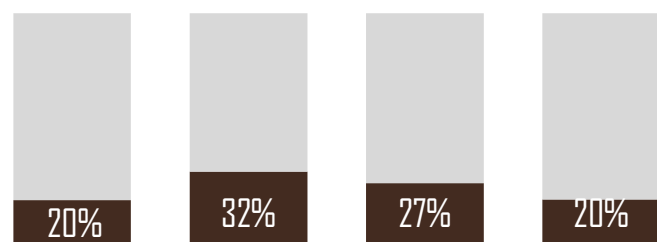




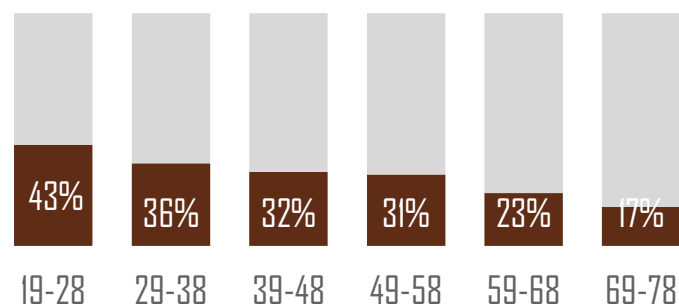
	Churn	Risk ind
15001	8%	Low
85062	31%	High

	Churn	Risk ind
Phoenix	22%	Low
Lafayette	36%	High

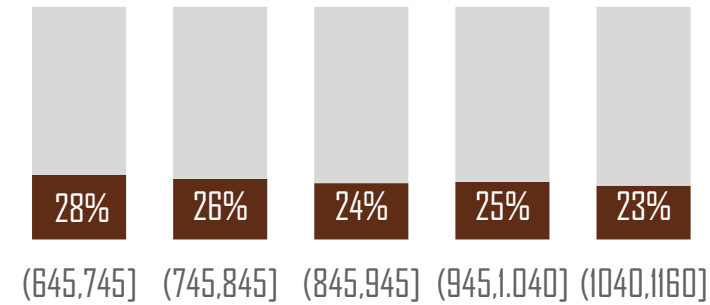
Bivariate analysis gives an indication about some possible characteristics of churn behavior



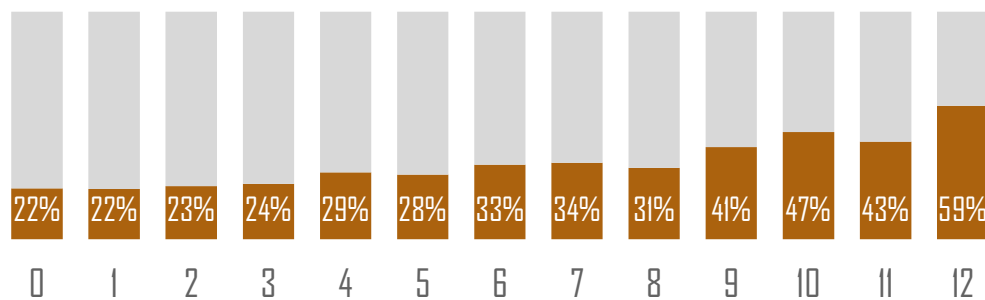
Churn rate vs Year



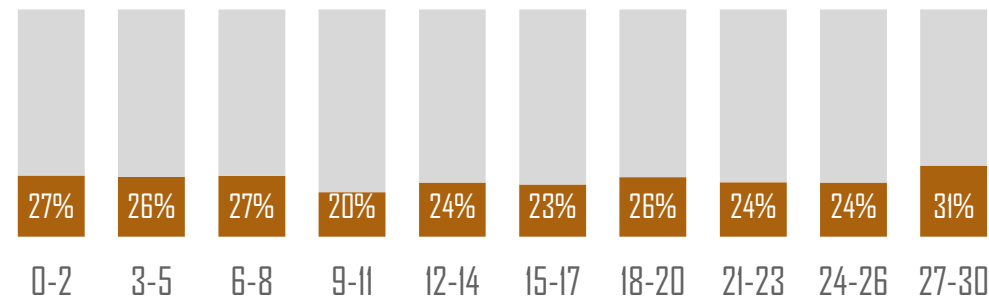
Churn rate vs Age



Churn rate vs Premium



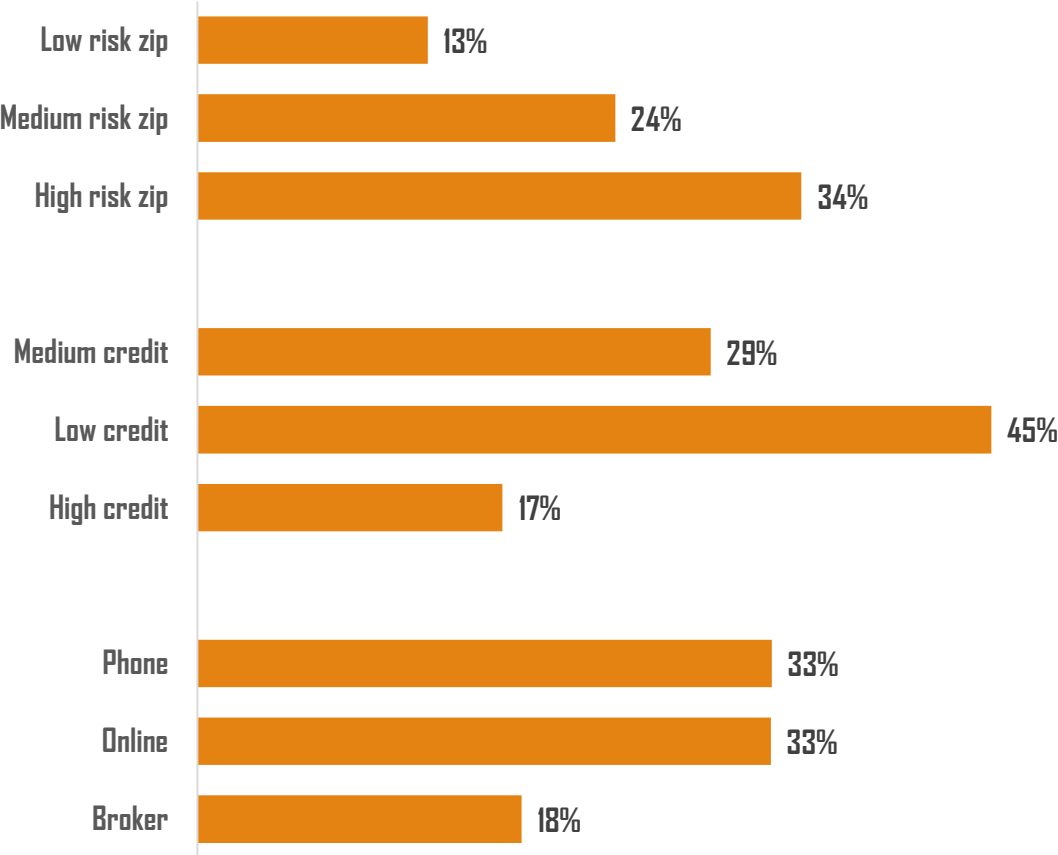
Churn rate vs # of Children



Churn rate vs Tenure

- Age and # of children seem to show linear indication to churn
- Churn rate is different for different years. There may be some thing different happening in years 2014,2015

For logistic problems, categorical variables are said to have highest predictive power.
To access variable importance and validate hypothesis we use weight of evidence

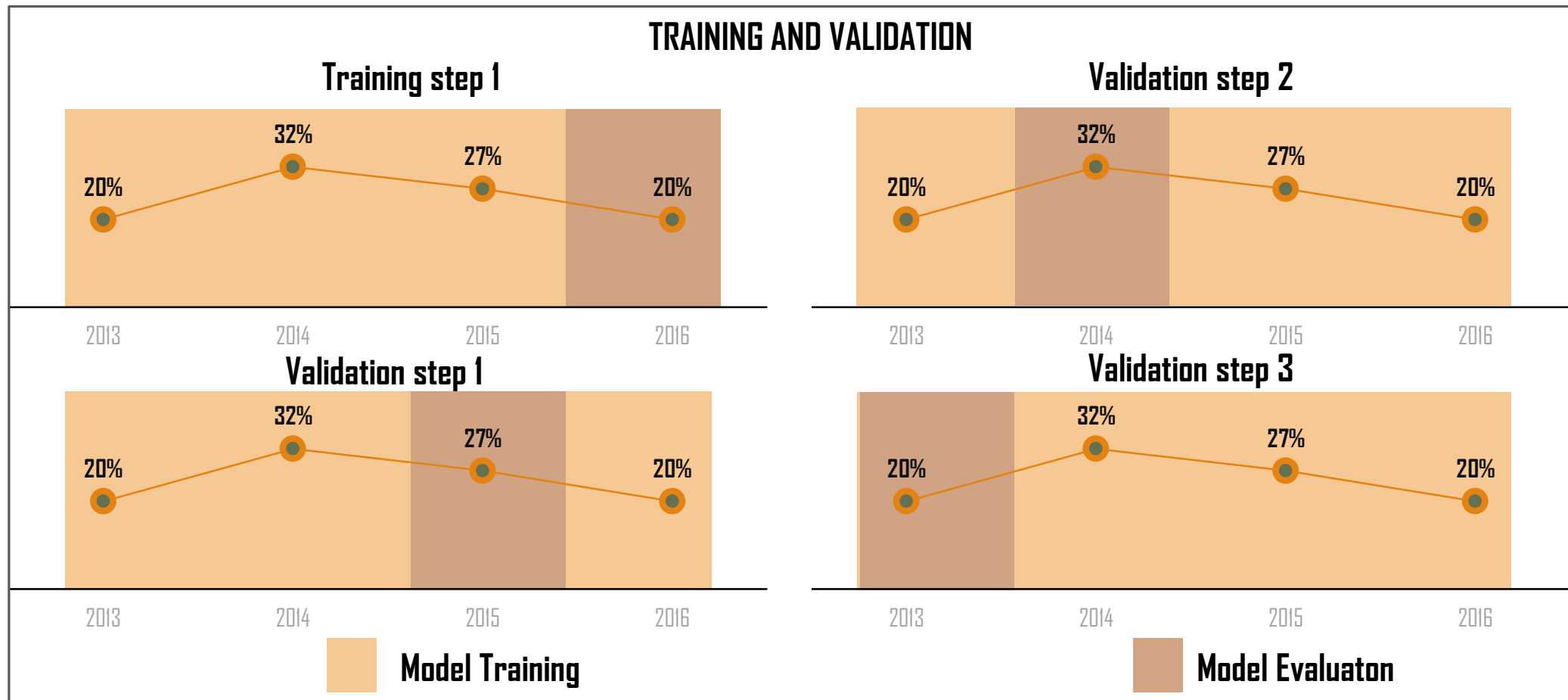


Variable	Variable importance
High Credit rating	0.21
Low Credit rating	0.21
Zip 5 mile risk rate	0.19
Broker	0.16
Low.Risk.City	0.12
High.Zip	0.11
Low.Zip	0.11
Phone	0.09
Low_risk_5mile	0.09
High.Risk.City	0.07
No of children	0.06
High_risk_5mile	0.06
PA state	0.05
VA state	0.04
X2016	0.03

- Credit history, the type of sales channel and the zip code churn risk seem to be good indicators
- The variable importance further validates the hypothesis that we have formulated

Some assumptions that were imposed while building the logistic model and these shaped our training testing and validation routine

- The predictors have no influence due to the year, i.e. independent of year, all predictors must be statistically significant in determining churn
- The Zip code risk score that is formulated has remained constant over Kangaroo's time of operation
- The data seen is a year snapshot of customer policies. The same customer does not appear in subsequent years.



Logistic regression was used to evaluate the probability of churn and the results were as follows

MODEL RESULTS

	Estimate	Std. Error	z value	Pr(> z)
Intercept	-1.34	0.20	-6.72	0.00
High credit rating	-0.69	0.10	-7.10	0.00
Low Credit rating	0.75	0.11	6.55	0.00
Broker sale	-0.88	0.08	-10.86	< .000002
# of Children	0.01	0.00	7.58	0.00
Low risk zip location	-0.33	0.13	-2.47	0.01
High risk location	0.32	0.10	3.21	0.00
Location 5 mile Churn rate	3.93	0.63	6.23	0.00
Purchase in 2016	-0.34	0.08	-4.24	0.00
Age of the holder	0.00	0.00	-2.07	0.04

* Numeric variables are squared

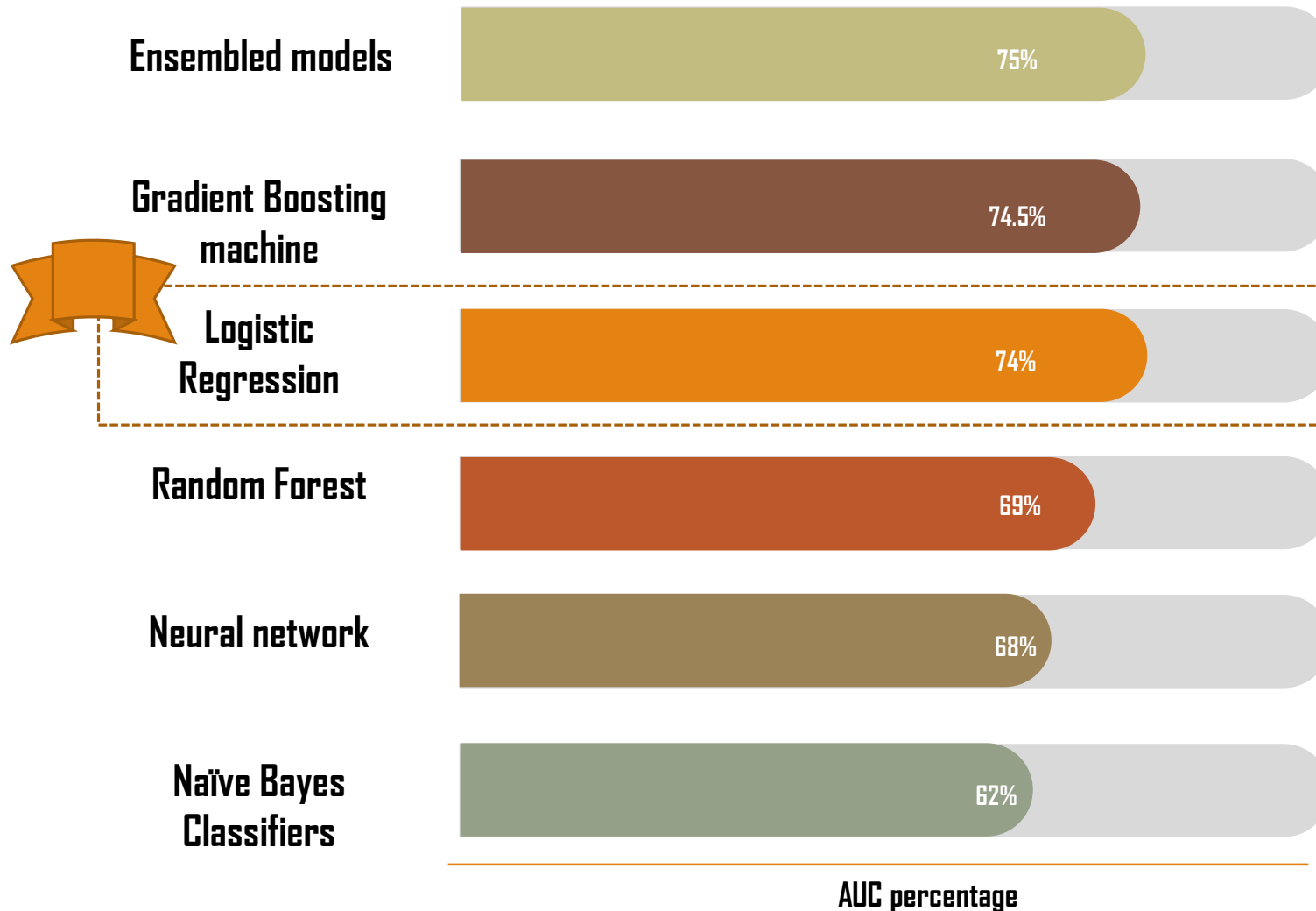
Confusion Matrix

	Predict Churn	Predict Active
Actual Churn	83	1201
Actual Active	47	4102

* Cutoff of 50% probability taken as churn

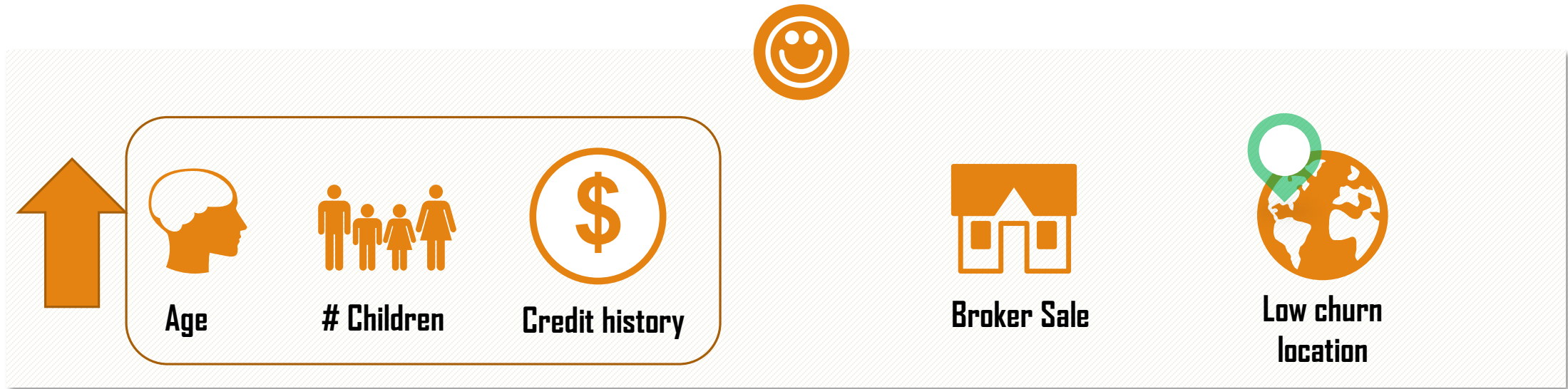
Using 2016 as validation dataset		Shuffling with years	
Misclassification rate	23%	Change in misclassification	+/- 1%
AUC Statistic	0.74	Change in AUC statistic	+/- 2%

Model fine tuning was done by considering additional modelling techniques and ensembling different models together



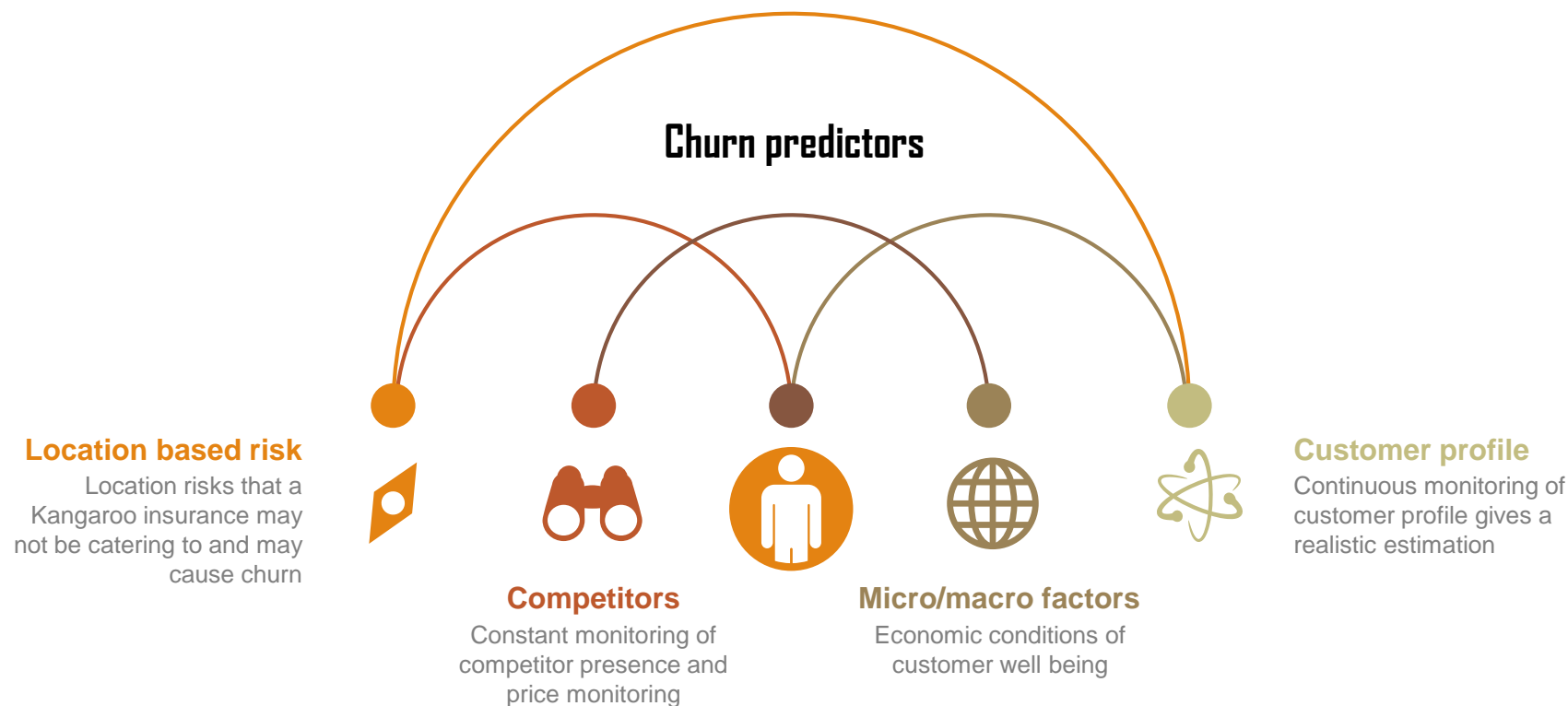
- Alternate models whose input was considered for ensembling
 1. Support Vector Machines
 2. KNN algorithm
 3. ADABOOST/XGBOOST models
- The best model that can be formulated is an ensemble model and that is very comparative to the performance of a logistic model
- **Selecting a logistic model has many advantages**
 1. It can give a under-the-hood look of which variables are important and what is its contribution to risk
 2. It is more stable provided the Out of time validation dataset follows same distribution of training and validation dataset

Based on the model building exercise, a specific set of profiles have been constructed for a churned customer and non churned customer



Next steps for better model accuracy statistics and having a detailed customer profile for churn

- While customer specific attributes are good indicators of churn, they represent only a part of the puzzle
- Customer experience with the company will be a firm determiner whether he will churn or not
- Misclassification rate is high, and tendency to predict churn isn't there. Therefore we need strong indicators of churn



Thank You