Задание:

Инструменты: для решения данного кейса я буду использовать язык R, Excel. Алгоритм «Random forest»

Первый этап.

На первом этапе проанализируем данные. Всего у нас имеется 50001 экспериментов. Для каждого из экспериментов у нас имеются результаты для 18 переменных.

Сразу отформатируем наши данные. Приведем данные из переменной «order\_date» в числовой вид, нам это потребуется для работы в R. Также изменим значения NA. Для того чтобы не произошло смещения NA мы меняем на -9999. Данные готовы для анализа.

Второй этап.

Предварительно, построив модель по всем переменным, убедимся, что переменные которые мы не будем включать в модель (по заданию кейса) не будут ключевыми. Так оно и оказалось, поэтому смело из наших данных мы убираем эти переменные и оставляем только 8 переменных.

Теперь у нас имеется отформатированная выборка из 8 переменных и 50001 наблюдения.

Третий этап.

На третьем этапе мы строим модель по уже отформатированной выборке.

Разбиваем данные на две выборки на «тренировочную выборку» и « тестовую выборку» в соотношении 3/4, 1/4 .

Как мы видим, наша модель сбалансирована. В дальнейшем это поможет нам в выборе метрики.



Строим «лес». Ограничиваем модель 500 -ми деревьями ( нам потребуется даже меньше, т. к. у нас не такая большая выборка).

Как мы и думали, нам понадобилось только 150 деревьев.

Четвертый этап.

На четвертом этапе мы получаем результаты анализа.

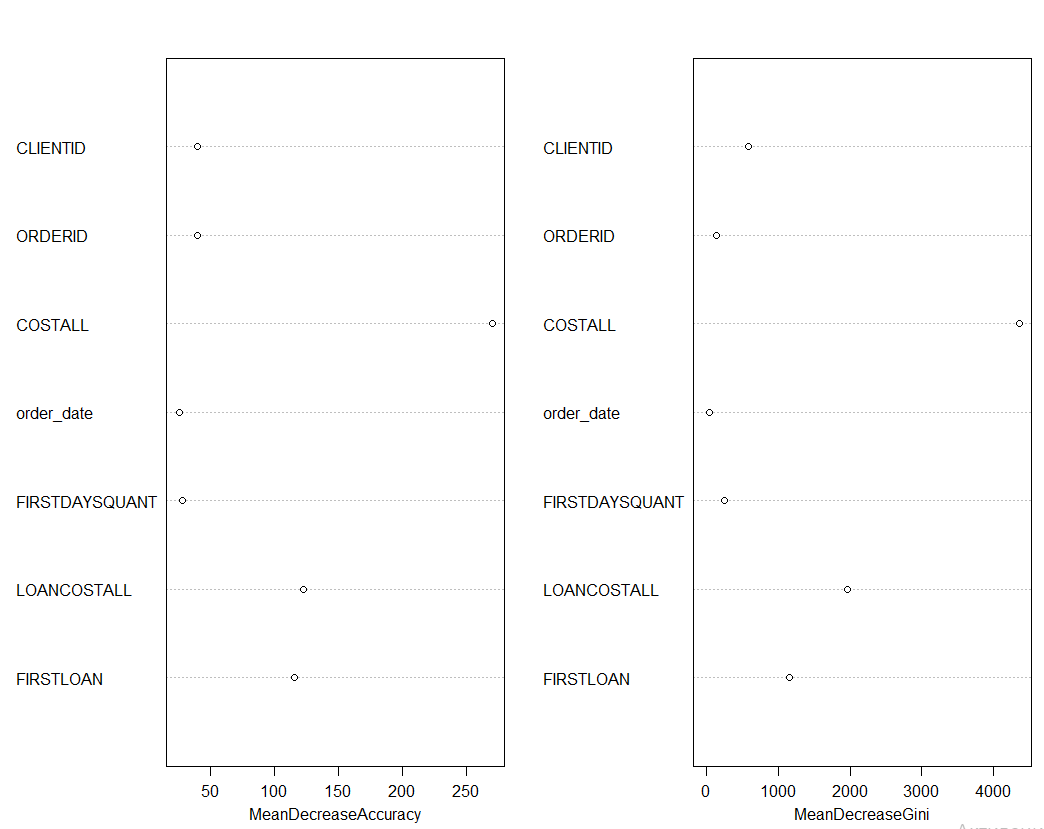
Прогоняем нашу обучающую и тестовую выборки.



Пятый этап.

На данном этапе мы будем анализировать наши данные.

На основании модели мы можем говорить о важности наших переменных для построения модели.



Как мы можем видеть самой главной переменной является «COSTAL», что же касается переменной «order\_date» ты мы ее можем безболезненно исключить из нашей модели.   
Для проведения анализа точности модели я буду использовать модель «F1». Так как данная метрика включает в себя две оценки «полноту» и «точность» ее результаты точнее, чем у данных метрик по-отдельности. Этой точности нам будет достаточно, да и данный критерий может выступать в качестве единственного критерия качества.



Значение «F-мера» для обучающей выборки 0,991, для тестовой выборки 0,991. Максимально возможный показатель равен 1. Исходя из полученных показателей мы можем утверждать, что наша модель имеет высокое качество.

Итоги.

Мы получили модель, построенную по алгоритму «Random Forest». Проанализировав точность данной модели с помощью метрики «Fмера», мы получили отличные показатели как для обучающей выборки (0,991) так и для тестовой (0,991).