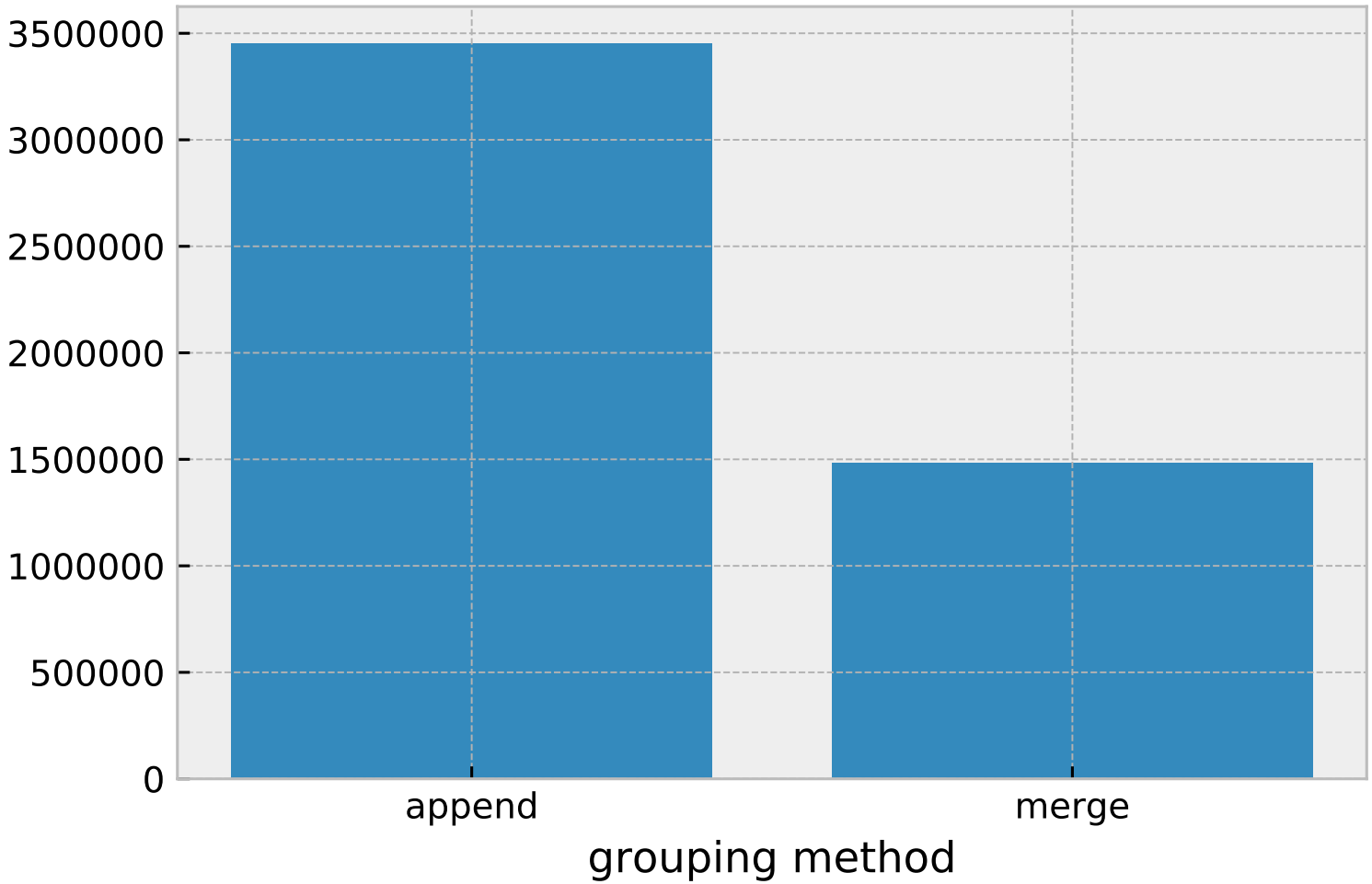
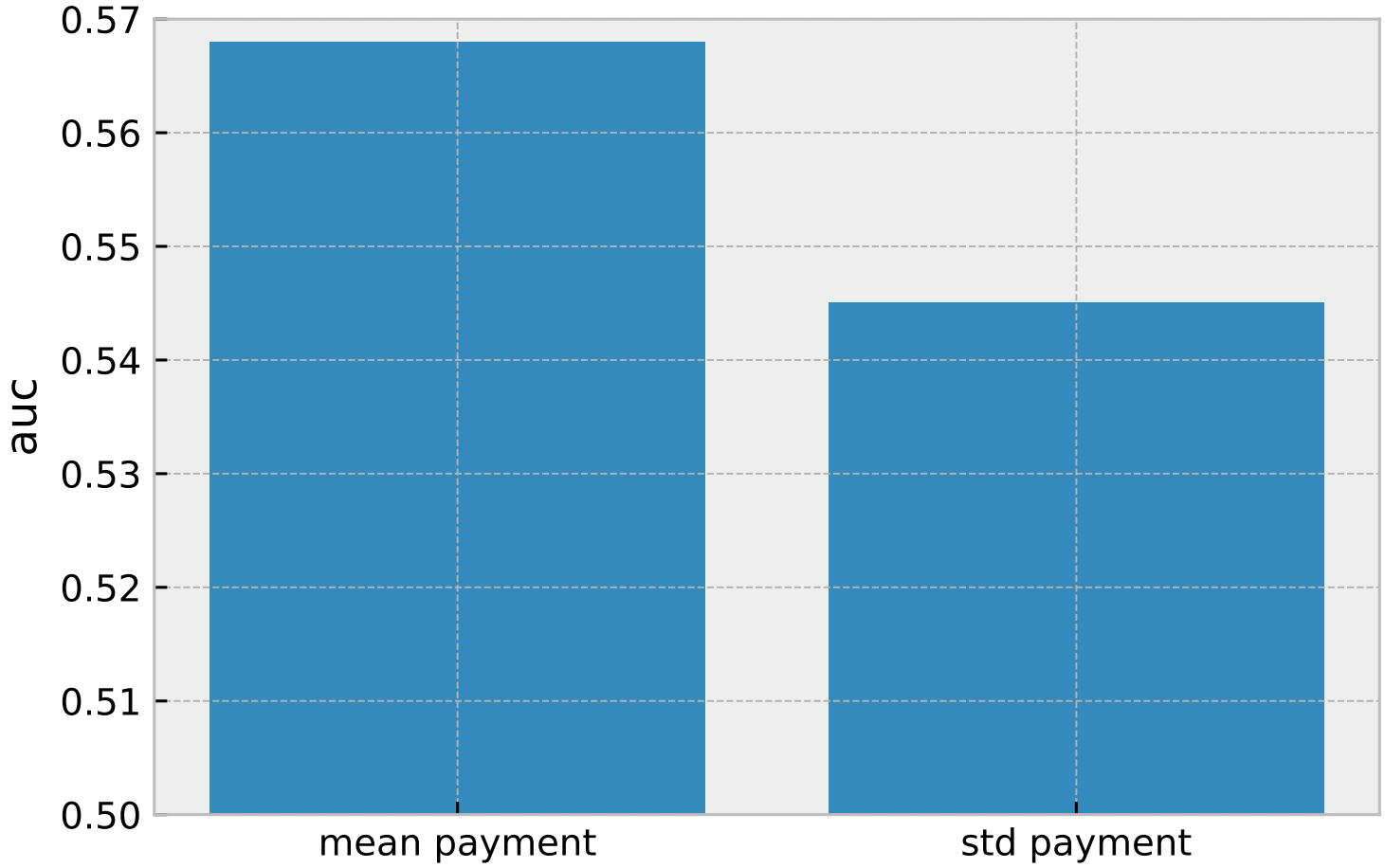
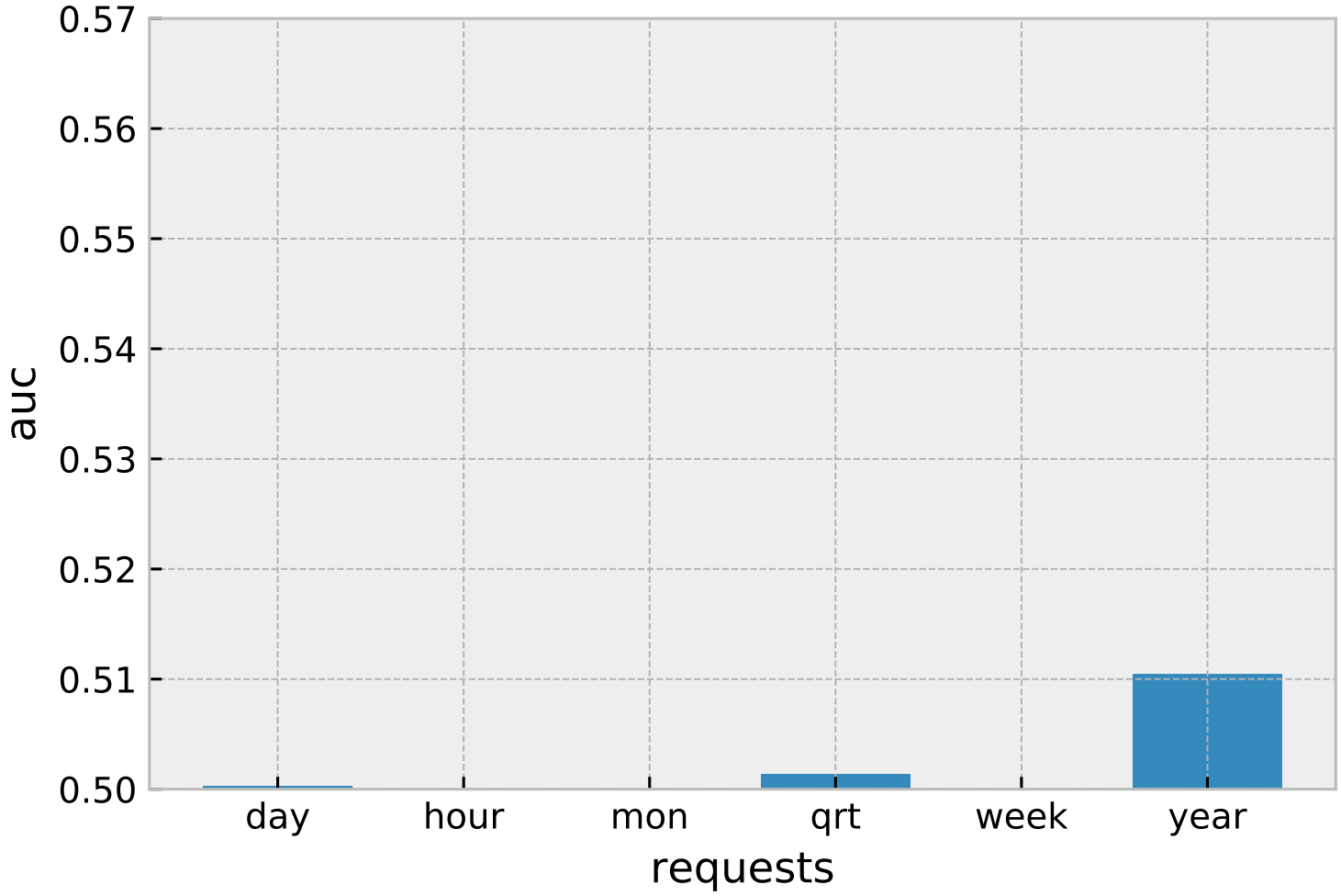


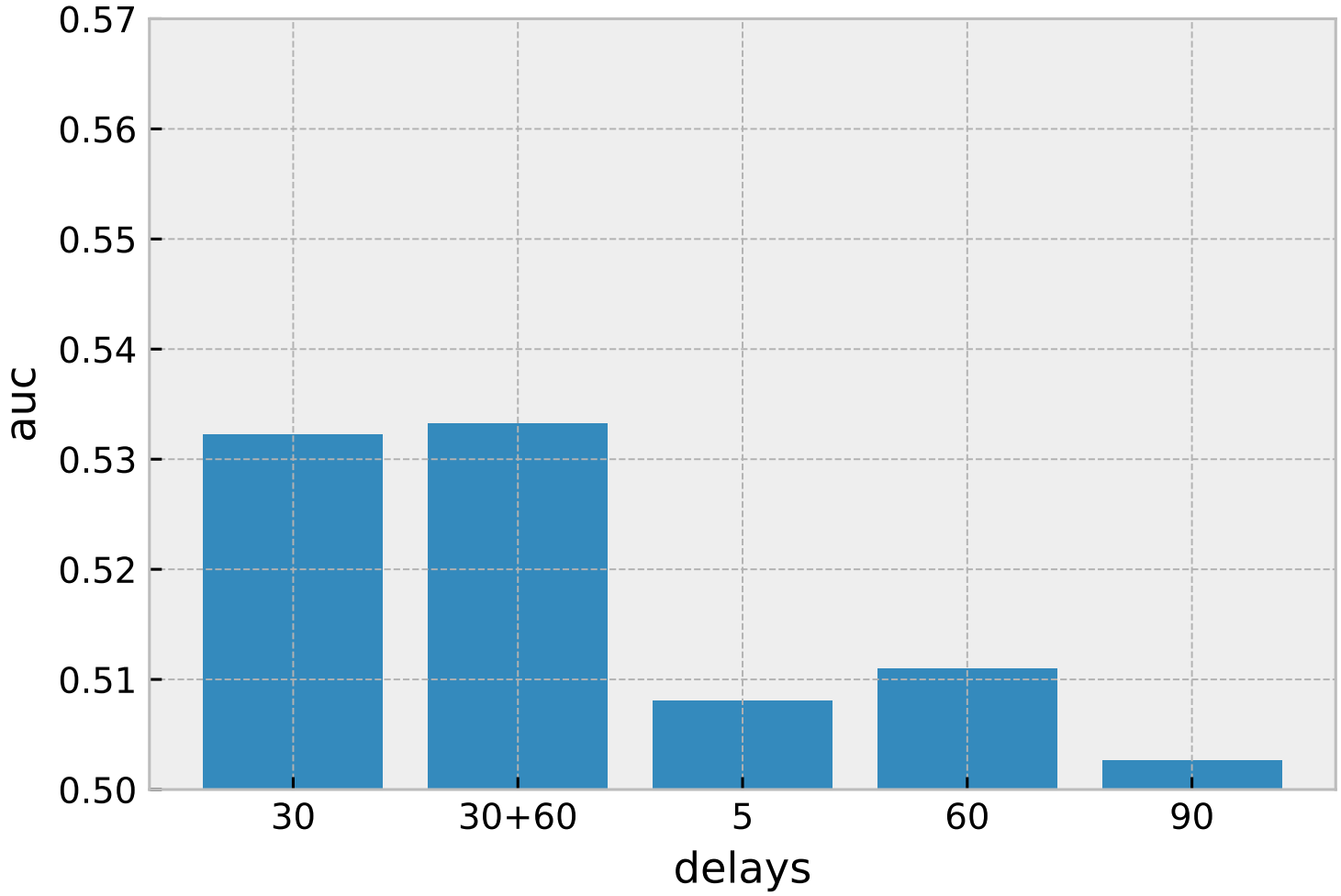
Данные из нескольких источников объединим в одну выборку посредством слияния. Оставим только уникальные элементы с заполнением недостающей информации по дубликатам.



Признаки, вносящие значимый вклад в auc возьмём за основу. Около 0.56 и 0.54 для среднего и стандартного отклонения в признаке длины записи 'text payment discipline'.



Исследуем важность группы признаков – количество запросов по временным промежуткам 'amt req *'. Значимым можно считать только признак 'amt req source year', на нём auc равен 0.51.



Исследуем важность группы признаков – число просроченных дней 'credit delay*'. Наиболее значимым является признак 'credit delay30' + 'credit delay60', на нём auc равен 0.53. Остальные группировки дают меньшее качество.

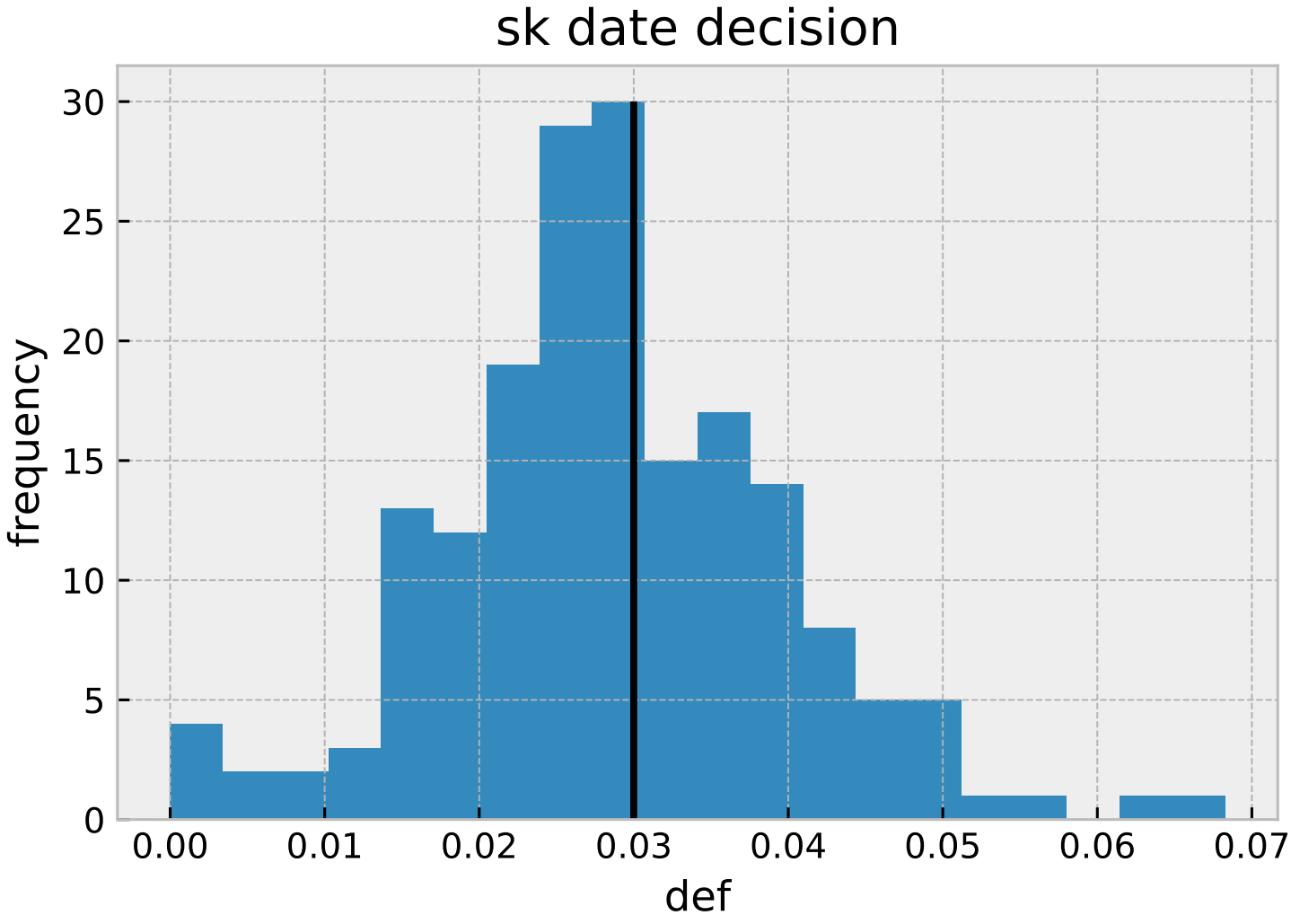
def credit currency	0	1
0	748510	22998
1	979	30

$AUC = 0.50017$

Попарной упорядоченности между признаком валюты и целевым вектором нет .

sk date decision	id	sk date decision	id
20160225.0	983	**20160301.0**	999
20160226.0	980	**20160302.0**	892
20160227.0	793	**20160303.0**	917
20160228.0	199	**20160304.0**	896
20160229.0	1075	**20160305.0**	630

Последние записи обучающей и первые записи контрольной выборок, упорядченных по номеру 'sk date decision'. Пересечения в выборках по данному признаку нет.



Однако распределение невыплаты по дате распределения заявки не вырождено и несёт в себе некоторую информацию. В этом должен быть смысл. На сайте конкурса об это признаке сказано отдельно.