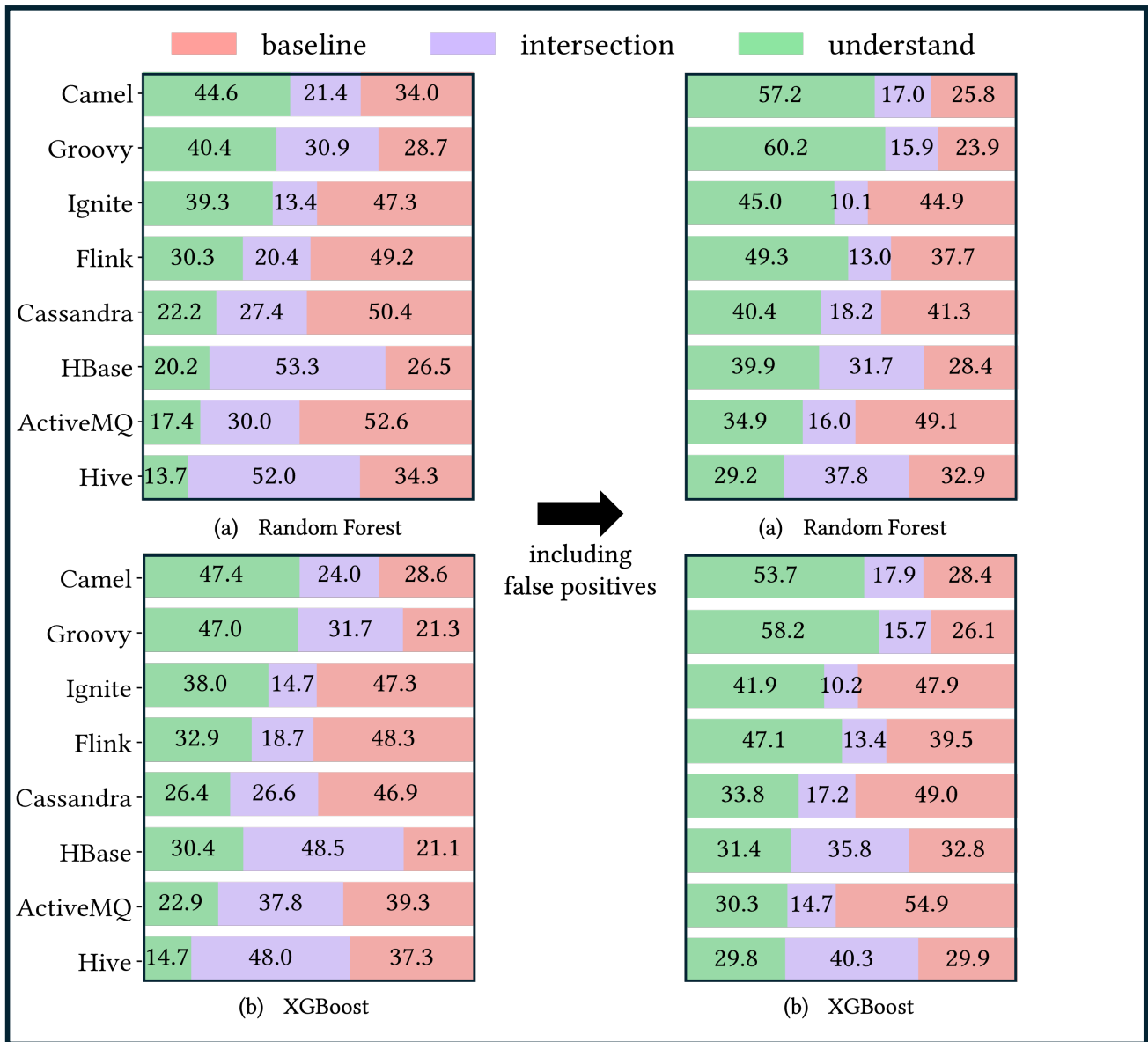


# Q3-1



**Set relationships after including false positives**

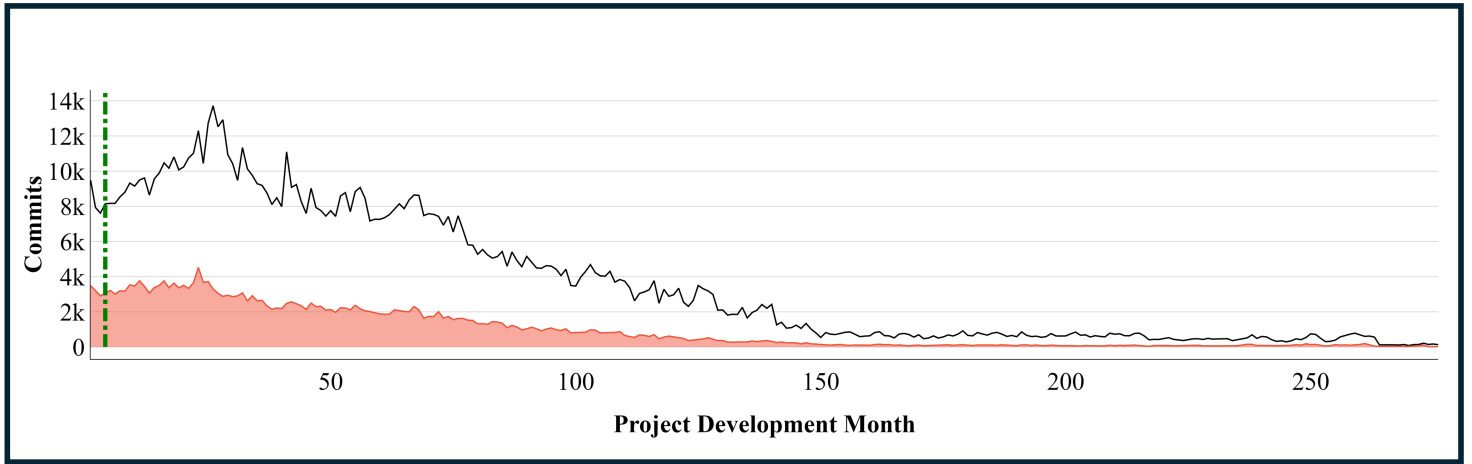
## Q3-2

$Wp$ : p-values from Wilcoxon rank-sum test,  $\delta$ : Cliff's delta effect size (l>m>s)

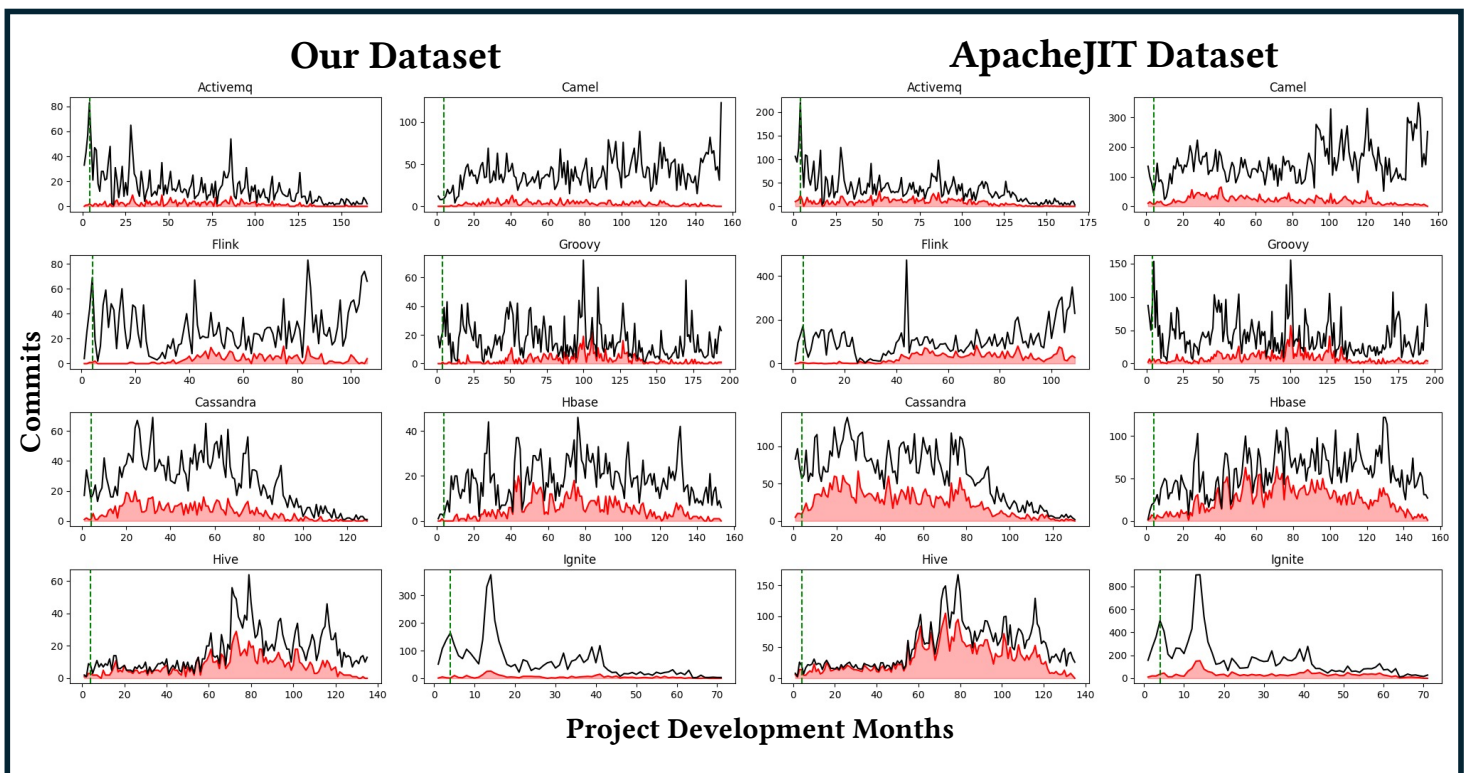
Project	Features	Random Forest				XGBoost			
		FPR	$Wp, \delta$	AUC	$Wp, \delta$	FPR	$Wp, \delta$	AUC	$Wp, \delta$
ActiveMQ	baseline	0.157	0.000, l	<b>0.573</b>	0.007, l	0.178	0.000, l	<b>0.561</b>	0.045, m
	combined	<b>0.087</b>		0.553		<b>0.131</b>		0.550	
Camel	baseline	0.061	0.000, m	0.541	0.725	0.070	0.000, l	0.539	0.279
	combined	<b>0.041</b>		0.542		<b>0.055</b>		0.529	
Flink	baseline	0.108	0.011, m	0.601	0.262	0.123	0.871	0.599	0.168
	combined	<b>0.093</b>		0.590		0.123		0.589	
Groovy	baseline	0.111	0.088	0.638	0.000, l	0.112	0.291	0.606	0.000, l
	combined	0.101		<b>0.698</b>		0.118		<b>0.693</b>	
Cassandra	baseline	0.194	0.000, l	0.633	0.001, l	0.193	0.000, l	0.614	0.001, l
	combined	<b>0.142</b>		<b>0.659</b>		<b>0.155</b>		<b>0.644</b>	
HBase	baseline	0.237	0.208	0.632	0.001, l	0.224	0.123	0.608	0.004, l
	combined	0.227		<b>0.646</b>		0.235		<b>0.627</b>	
Hive	baseline	0.394	0.172	0.692	0.291	0.419	0.068	0.674	0.725
	combined	0.387		0.697		0.408		0.677	
Ignite	baseline	0.097	0.002, l	0.557	0.425	0.095	0.039, m	0.563	0.402
	combined	<b>0.076</b>		0.561		<b>0.077</b>		0.568	

**Performance comparison in terms of FPR and AUC**

# Q5



The distributions of defect-inducing commits over time in the reference



The distributions of defect-inducing commits over time in our study